# Advanced Motif Analysis on Text Induced Graphs

Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation von Thomas Otmar Arnold aus Lindenfels
Tag der Einreichung: 19. April 2018, Tag der Prüfung: 24. Mai 2018
Darmstadt — D 17

1. Gutachten: Prof. Dr. Karsten Weihe
2. Gutachten: Prof. Dr. Iryna Gurevych
3. Gutachten: Prof. Dr. Matthias Müller-Hannemann

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Informatik
Fachgebiet Algorithmik

Advanced Motif Analysis on Text Induced Graphs

Genehmigte Dissertation von Thomas Otmar Arnold aus Lindenfels

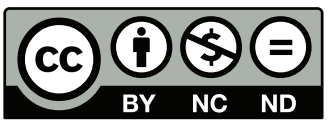
1. Gutachten: Prof. Dr. Karsten Weihe
2. Gutachten: Prof. Dr. Iryna Gurevych
3. Gutachten: Prof. Dr. Matthias Müller-Hannemann

Tag der Einreichung: 19. April 2018
Tag der Prüfung: 24. Mai 2018

Darmstadt — D 17

Bitte zitieren Sie dieses Dokument als:
URN: urn:nbn:de:tuda-tuprints-74428
URL: http://tuprints.ulb.tu-darmstadt.de/7442

Dieses Dokument wird bereitgestellt von tuprints,
E-Publishing-Service der TU Darmstadt
http://tuprints.ulb.tu-darmstadt.de
tuprints@ulb.tu-darmstadt.de



# Erklärung zur Dissertation

Hiermit versichere ich, die vorliegende Dissertation ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 19.04.2018

_____

(Thomas Otmar Arnold)

# Abstract

Motif analysis counts the number of reoccurring patterns (or *motifs*) in a graph and connects these statistical numbers to the intrinsic semantics of the graph. In this thesis, we will demonstrate the potential of motif analysis on textual data, and introduce new concepts that extend conventional motifs. In particular, we will focus on three main research questions:

1. Can we use graph motifs to assess text quality?

Based on the open encyclopedia Wikipedia, we transform articles of various quality levels into graph structures. There, we find motifs that indicate high or low article quality, and we connect these motifs to linguistic patterns. We also show that a qualitative analysis of the most relevant patterns can yield fruitful insights to our understanding of quality. We then take a look at quality from a very different angle and analyze motifs in the user interaction of collaborative writing communities. These interaction motifs allow us to assess the overall online community success, measured by a combination of growth and user traffic. Certain combinations of user groups show consistent beneficial or detrimental effects on the community performance.

2. How do motifs change over time?

Having established that motif analysis can detect quality on different levels, we now focus at the progression of motifs in dynamic graphs. We take another look at Wikipedia articles, in particular at local text changes in article revisions. To capture patterns in these text revisions, we introduce *metamotifs*, or motifs of motifs. We also define the novel concept of motif stability - motifs of high stability tend to persist in dynamic graphs, motifs of low stability almost always get changed into other motifs. We present strong correlations between motif stability, established motif characteristics and the quality of the source text.

3. Are metamotifs (motifs of motifs) an improvement over simple motifs and methods?

Finally, we confirm the capabilities of metamotifs, but also quantify their predictive power in a classification experiment of political speeches. To generalize from surface text level, we use semantic frames, which are more abstract than words. With a combination of semantic frames and metamotif analysis on US presidency and German Bundestag data, we confirm that metamotifs outperform traditional motifs and simpler approaches when used as machine learning features.

# Zusammenfassung

Motivanalyse zählt die Anzahl von wiederkehrenden Mustern (auch *Motive* genannt) in einem Graphen und setzt diese statistischen Zahlen mit der intrinsischen Semantik des Graphen in Verbindung. In dieser Arbeit werden wir das Potenzial von Motivanalyse in Textdaten aufzeigen und neue Konzepte vorstellen, die konventionelle Motive erweitern. Insbesondere werden wir uns auf drei Hauptforschungsfragen konzentrieren:

1. Können Graphen-Motive zur Beurteilung von Textqualität verwendet werden?

Basierend auf der freien Online-Enzyklopädie Wikipedia transformieren wir Artikel verschiedener Qualitätsstufen in Graphstrukturen. Dort finden wir Motive, die auf hohe oder niedrige Artikelqualität hinweisen, und bringen diese Motive mit linguistischen Mustern in Verbindung. Anhand einer qualitativen Analyse der relevantesten Muster demonstrieren wir, dass Motive neue Erkenntnisse für unser Verständnis von Qualität liefern können. Dann betrachten wir Qualität aus einem ganz anderen Blickwinkel und analysieren Motive in der Interaktion von Autoren in kollaborativen Schreibprozessen. Diese Interaktionsmotive ermöglichen es uns, die Gesamtleistung einzelner Online-Gemeinschaften zu bewerten, gemessen an einer Kombination aus inhaltlichem Wachstum und Nutzeraktivität. Dabei zeigen bestimmte Kombinationen von Benutzergruppen konsistente positive oder negative Auswirkungen auf den Erfolg der Gemeinschaft.

2. Wie verändern sich Motive im Laufe der Zeit?

Nachdem wir gezeigt haben, dass Motivanalyse die Qualität von Text auf verschiedenen Ebenen erkennen kann, konzentrieren wir uns nun auf die Veränderung von Motiven in dynamischen Graphen. Wir werfen dafür einen weiteren Blick auf Wikipedia-Artikel, insbesondere auf lokale Textänderungen in Artikelrevisionen. Zur Erkennung von Motiven in diesen Textrevisionen führen wir *Metamotive* - Motive von Motiven - ein. Zusätzlich definieren wir ein neuartiges Merkmal von Motiven, genannt Motivstabilität. Motive hoher Stabilität bleiben in dynamischen Graphen meist bestehen, während sich Motive niedriger Stabilität eher in andere Motive verwandeln. Dieses Merkmal ermöglicht es uns, starke Zusammenhänge zwischen bestimmten strukturellen Eigenschaften von Motiven, und ihrer Erwünschtheit in Bezug auf Textqualität herzustellen.

3. Sind Metamotive (Motive von Motiven) eine Verbesserung gegenüber einfachen Motiven und Methoden?

Zum Abschluss dieser Arbeit bestätigen wir die Mächtigkeit von Metamotiven und messen ihre Vorhersagekraft in einem Klassifikations-Experiment von politischen Reden. Anstatt Motive direkt im Text zu betrachten, verwenden wir semantische Rahmen (semantic frames) als Abstraktionsebene. So kombinieren wir semantische Rahmen und Metamotiv-Analysen, um Texte von US-Präsidentschafts-Kandidaten und Debatten des deutschen Bundestags zu analysieren. Durch verschiedene maschinelle Lern-Experimente bestätigen wir, dass Metamotive eine höhere Trennschärfe besitzen als traditionelle Motive und einfachere Ansätze.

# Contents

# 1 Introduction

Even in our modern times of famous YouTube stars and video messaging, textual resources remain an extremely important source of information with undeniable benefits over other forms of knowledge media. For instance, online search engines are dependent on the convenient searching capabilities of text data, and the adaptability and flexibility of text enables collaborative knowledge bases like Wikipedia. In the scientific world, text is the established medium of preserving and sharing knowledge. Regardless of media type, we can be overwhelmed by the huge amount of easily available data. As a consequence, there is a growing need for helpful tools and automatic ways to search, filter or consume information that fits our needs. Therefore, the area of natural-language processing (NLP) is concerned with processing huge amounts of human language data. Common NLP tasks are closely connected to many areas of our life. Text can be automatically translated to a different language, search engines provide short summaries of relevant text sources, and speech recognition software allows us to communicate with intelligent personal assistants, like Siri or Amazon Echo. One major issue in all NLP tasks is the quality of the output. Translated text is only useful if it still conveys the content and intention of the source material, and a good summary should not only contain the most important facts, but also present them in a understandable and comprehensible format.

The NLP research community has utilized a variety of techniques and approaches, including various forms of machine learning, and many ways of representing textual data. Graph representations are a natural way to model text and other structured data. These

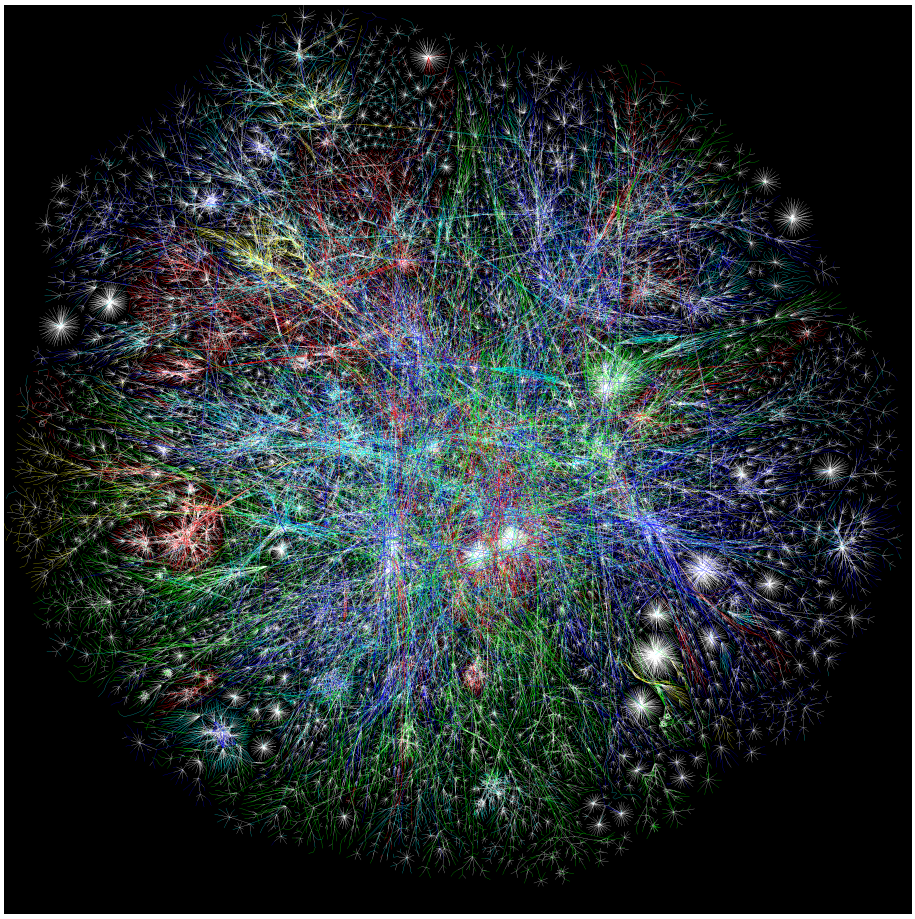


**Figure 1.1:** A graph representation of the internet of November 2003. Each online server is represented as a node. Each connection between two servers is shown as a colored edge. The colors show servers and connections from different parts of the world. [69]

graphs, or networks, project units of data to graph nodes. If two units of data are connected with a pre-defined relation, the two corresponding nodes are also connected in the graph represenation. This form of data representation can be used to simply visualize complex data structures. Figure 1.1 shows an illustration of the World Wide Web of 2003, which has been used to teach students about the internet and its growth in specific areas [69]. In addition to these possibilities of visualization, graph structures also enable specialized graph analysis and algorithms, which can lead to deeper knowledge on the source data.

In this thesis, we explore graphs that represent textual data. In particular, we are interested in reoccurring patterns ("motifs") in these graph models. Language is naturally structured, and often follows syntactic patterns, like subject-verb-object (SVO). Using different graph representations for texts, the resulting graphs can contain motifs of very different nature. These motifs can allow us to draw conclusion about hidden patterns in our language, and also their consequences and implications. We especially want to focus on the connections between motifs and text quality. As mentioned, detecting and creating high quality text is an important task in many NLP applications. If we analyze patterns that naturally occur in our language, we may be able to improve our understanding of text quality. This knowledge is essential to create better solutions and tools for general human-computer interaction.

In this thesis, we present different applications of graph motif based algorithms on various textual data sets. In addition, we also explore extensions of established approaches, discuss motifs in graphs that change over time, and evaluate the potential of novel types of motifs.

## 1.1 Structure of this Thesis

First, we will look at the core research questions of this thesis in Chapter 2, and cover the theoretical foundations of graph theory and graph motifs in Chapter 3. Then we present two motif analysis experiments on the subject of quality in text based data, but with very different approaches: In Chapter 4, we directly discover motifs in encyclopedic Wikipedia articles that are correlated with high or low article quality. We then remain in the context of encyclopedic online platforms, but investigate motifs in a very different graph representation. Chapter 5 presents motifs of user interaction in the collaborative

writing platform Wikia. We will demonstrate that some motifs have positive or negative implications on the success of the writing community. Having discussed the overall usefulness of motifs in textual data, Chapter 6 draws the focus on the evolution of motifs over time, and introduces the notion of metamotifs: motifs of motifs. We quantify the predictive power of metamotifs in Chapter 7 in the setting of political speeches, and their interpretability. Finally, Chapter 8 draws a summarizing conclusion, including possible extensions.

# 2 Research Questions

Graph- or network-based approaches have been successfully applied in different scientific disciplines, like biology, biochemistry or electrical engineering. In particular, searching and analyzing recurrent substructures, or motifs, of these graphs has led to very interesting insights. In my doctoral research, I applied and also extended the concept of motif analysis and its uses on text based graphs. The experiments and evaluations in this thesis discuss three main research questions:

1. Can we use graph motifs to assess text quality?
2. How do motifs change over time?
3. Are metamotifs (motifs of motifs) an improvement over simple motifs and methods?

---

Can we use graph motifs to assess text quality?

---

What is a high quality text? This question is both generic and impossible to answer. The requirements of a well-written factual summary, a compelling crime story or a recipe book are very different. But even when we only consider a very specific text type and purpose, it is hard to specify and define the spectrum of text quality. Some aspects are easier to grasp, like readability, text coherence or grammatical correctness. compared to more impalpable facets like creative use of language, but measuring them is still a difficult task. We hypothesize that text, as a structured medium, contains characteristic patterns that can be used to classify and understand many textual properties, including

quality. We want to use motif analysis on textual data to prove this hypothesis, and also investigate the interpretability of motifs. In addition to their usefulness as predictive features, distinctive language motifs might help us understand underlying reasons for their existence and prevalence.

### How do motifs change over time? What can we learn from motif changes?

Language is not static. It changes from century to century, from year to year, and in some instances, even from one hour to the next. If language is in a constant flow, so are its characteristics, attributes and patterns. Understanding how the patterns in our texts and speech change, and why, can derive a new dimension of knowledge about our language, and the people that use it. We want to explore and formalize locally changing textual motifs, and evaluate applications of this approach.

### Are metamotifs an improvement over simple motifs and other methods?

In previous research, the nature of graph motifs mostly followed the same definition. Motifs are often fixed to connected subgraphs of limited maximum size. However, investigating the interplay between motifs is usually neglected. We want to find motifs within the usage of textual motifs and evaluate the power of these *metamotifs* in exemplary experimental settings.

# 3 Theoretical Fundamentals

The main research questions of this thesis focus on exploration and extension of graph motif analysis on textual data. This chapter explains the necessary theoretical concepts and methods, and presents related work. We start this chapter with an introduction to general graph theory (see Section 3.1). Then, we will explain the concept of graph motifs with its applications and challenges (see Section 3.2).

## 3.1 Graph Theory

Formally, a graph or network is an ordered pair $G = (V, E)$, with a set of nodes or vertices $V$ and a set of edges $E \subseteq V \times V$. Graph can be directed or undirected. In an undirected graph, every edge is an unordered pair of nodes, since there is no direction associated with the edge. In a directed graph, every edge is an ordered pair of nodes, where the order represents the direction of the edge that links the two nodes. For instance, an undirected edge

$e \in E, e = \{v1, v2\}$ with $v1, v2 \in V$

connects the two nodes $v1$ and $v2$, but there is no defined start or end node of the edge. A directed edge

$e \in E, e = (v1, v2)$ with $v1, v2 \in V$

also connects the two nodes $v1$ and $v2$. In this case, $v1$ is called the start node of $e$, and $v2$ is called the end node of $e$. Also, $e$ is called an outgoing edge of $v1$ and an incoming edge of $v2$. The indegree and outdegree of a given node $v1 \in V$ is the number of

incoming and outgoing edges of $v1$, respectively. In both directed and undirected cases, $e$ is considered to be incident to both $v1$ and $v2$. Since $v1$ and $v2$ share an incident edge, $v1$ is considered adjacent to $v2$, and vice versa. The degree of a node is the number of edges incident to the node. Figures 3.1 and 3.2 show example visualizations of undirected and directed graphs. It is important to note that the visualization of a graph is not unique. Although a graph is unambiguously defined by its sets of nodes and edges, there are an infinite number of possible visualizations.

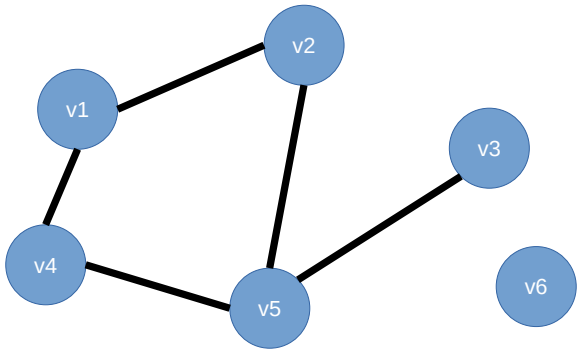



**Figure 3.1:** Example visualization of the undirected graph $G = (V, E)$ with node set $V = \{v1, v2, v3, v4, v5, v6\}$ and edge set $E = \{\{v1, v2\}, \{v1, v4\}, \{v2, v5\}, \{v3, v5\}, \{v4, v5\}\}$. Note that node $v6$ is part of the graph, although not connected to any other node.

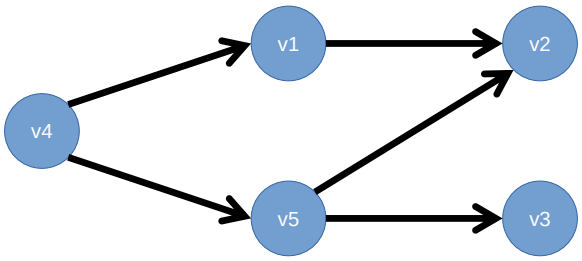



**Figure 3.2:** Example visualization of the directed graph $G = (V, E)$ with node set $V = \{v1, v2, v3, v4, v5\}$ and edge set $E = \{(v1, v2), (v4, v1), (v4, v5), (v5, v2), (v5, v3)\}$. In contrast to undirected graphs, edges are ordered instead of unordered pairs of nodes.

Graph models can be augmented by assigning weights to all edges. For this purpose, we define an edge cost function $c : E \to \mathbb{R}$ that maps each edge in the graph to a real number. These weighted graphs can model pairwise connections that have numerical values. For example, edges might represent roads of a road network, with edge weights representing the length of the road, or the time needed to travel from the start location to the destination.

Normally, multiple edges connecting the same set of two nodes (in the same direction, in the case of a directed graph) are not allowed. They are permitted in so called multi-graphs. There, the set of edges is not a subset, but a multiset of node pairs. In other words, node pairs may exist multiple times. See Figure 3.3 for an example visualization of a directed multigraph with edge weights.

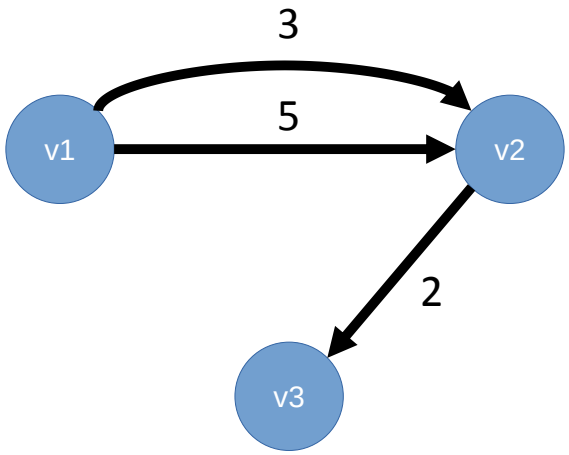



**Figure 3.3:** Example visualization of the directed multigraph $G = (V, E)$ with node set $V = \{v1, v2, v3\}$ and edge set $E = \{e1, e2, e3\}$, $e1 = (v1, v2)$, $e2 = (v1, v2)$, $e3 = (v2, v3)$. The cost function $c : E \to \mathbb{N}$ assigns the following integral values to all edges: $c(e1) = 3$, $c(e2) = 5$ and $c(e3) = 2$

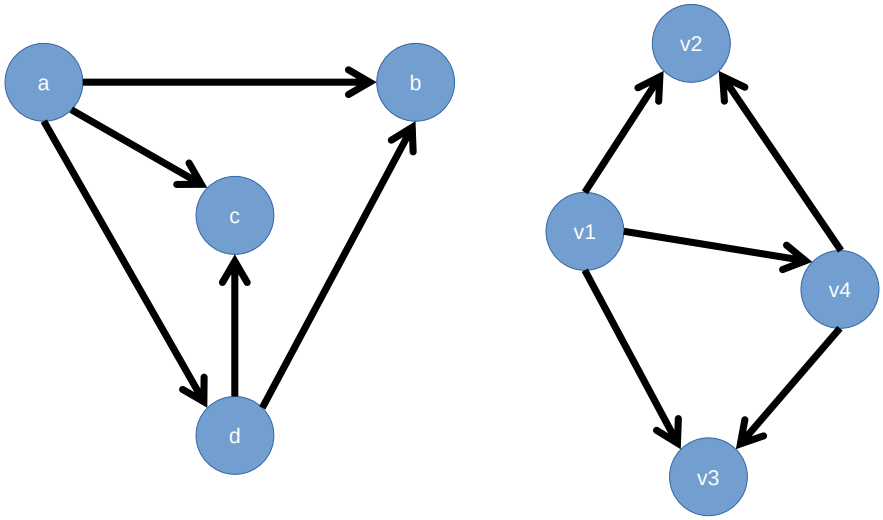



**Figure 3.4:** Example of graph isomorphism. These two graphs are isomorphic under this isomorphism $f : V_1 \to V_2$: $f(a) = v1$, $f(b) = v2$, $f(c) = v3$ and $f(d) = v4$.

Graph isomorphism is the final important concept of graph theory that we will need for the subject of motifs. An isomorphism of two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is a bijection between the node sets $V_1$ and $V_2$

$f : V_1 \to V_2$

so that two nodes $a, b \in V_1$ are adjacent in $G_1$ if and only if $f(a)$ and $f(b)$ are adjacent in $G_2$. Two graphs are isomorphic if an isomorphism exists between these two graphs. In other words, isomorphic graphs are structurally identical. The two graphs in Figure 3.4 are isomorphic, even though they look very different.

Graphs are used in many different scenarios and disciplines to model a variety of networks and relations. In computer science, graphs are often used to model communication networks, data organization, etc. For example, a directed graph can represent the link structure of a website, with web pages mapped to nodes and links from one page to another modeled with directed edges. Similar approaches were used in social media, biology, chemistry, travel business, computer chip design, and many other fields.

There is a large scientific body of methods and applications of graph analysis [2, 3]. Graph mining – the art of detecting and analyzing patterns and structures in graphs – is the specific focus of the surveys [31, 40].

It seems reasonable to classify graph analysis techniques by the level of granularity they address. Elementary statistical measures such as the node degree distribution operate on the level of single nodes and edges. In the opposite extreme case, on the global level, the structure of a graph is captured in a single (scalar) numerical value. Examples for global measures are the average shortest path length, the diameter, as well as simple characteristics such as node and edge count. See the above-mentioned surveys [31, 40] for a systematic discussion.

The first paper in the history of graph theory is assumed to be published in 1736 by Leonhard Euler on the Seven Bridges of Königsberg [20]. The goal in this problem is a path through Königsberg that crosses each of the seven bridges of the city exactly once. Euler reformulated and abstracted the problem by using a graph representation (Figure 3.5). He then observed that such a walk is possible if and only if the graph has exactly zero or two nodes with odd degree, since those have to be either start or end node of

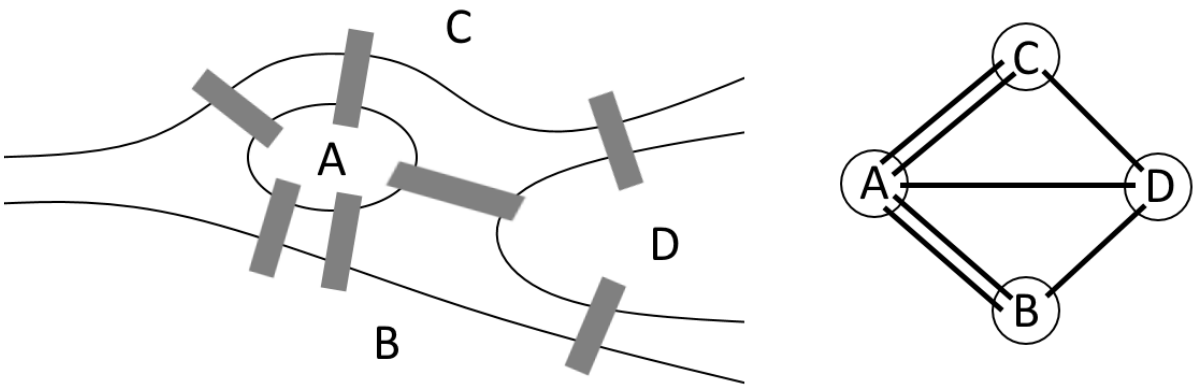



**Figure 3.5:** Schematic image and graph representation of the Seven Bridges of Königsberg problem. The core question of this problem: Is there a path that crosses each bridge exactly once?

the walk. In the problem of Königsberg, there are four nodes with odd degree, therefore Euler proved that this problem has no solution.

In the last years, graph-theoretic methods have proven particularly useful in natural language processing and linguistics, since textual data often fits well to discrete structure and models. For instance, constituency- or dependency-based parse trees follow tree-based structures, and map the syntax of natural language into a hierarchical graph. In computational linguistics, semantic networks that connect words to related words, have proven to be highly beneficial [88]. This usefulness has enabled several projects of high value for the natural language community, including WordNet [59] and VerbNet [75], or related projects like FrameNet [15]. Another prove for the fruitful combination of graph theory and computer linguistics is TextGraphs workshop series[1]. This annual workshop was introduced in 2006 as a platform to share knowledge about the application of graph-related methods to natural language challenges.

## 3.2 Graph Motifs and Motif Analysis

Many networks, including social, biological or chemical networks, can be represented as a graph. Every graph consists of its various subcomponents, or subgraphs. The goal of motif analysis is to extract and examine these subgraphs, and thereby derive new knowledge about the source data.

*Motifs* are basically small, connected subgraphs of a bigger network. They can vary in size, and are often limited to a fixed number of nodes, typically three or four. Figure 3.6 shows all motifs on three nodes in directed graphs, and an example graph with three highlighted motifs can be seen in Figure 3.7. The exact definition of motifs varies throughout the scientific literature. Some articles regard only subgraphs with excep-

[1] http://www.textgraphs.org/

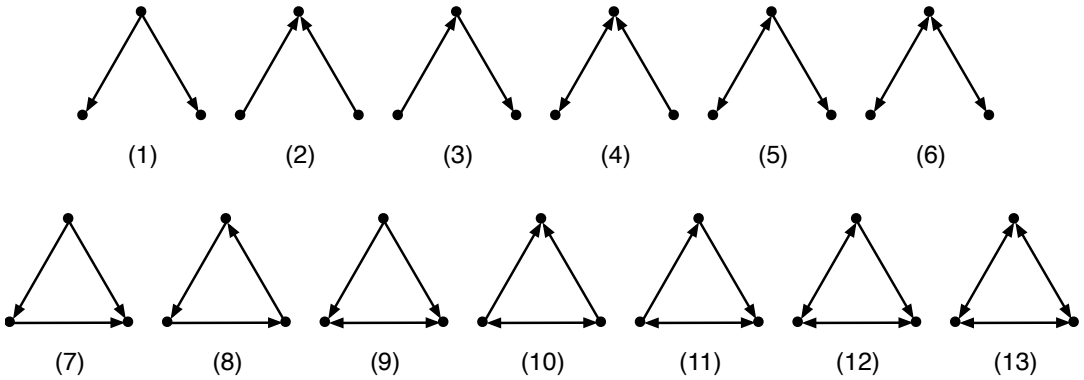



**Figure 3.6:** The directed motifs on three nodes. A double arrow indicates the presence of two mutually opposite arcs.

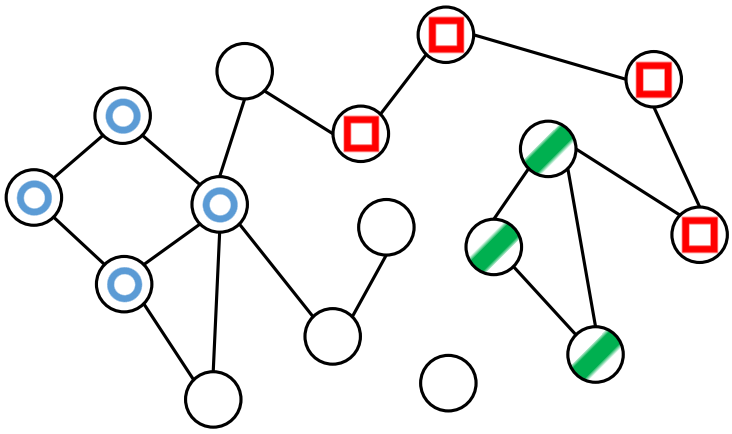


**Figure 3.7:** Example visualization of motifs in an undirected graph, with three highlighted graph motifs - two motifs with four nodes, and one motif with three nodes.

tionally high frequency to be motifs of the network, and subgraphs with exceptionally low frequency are called *anti-motifs* [16, 62]. Other articles use the term motif for all subgraphs, and search for most interesting motifs within all extracted ones [18, 73]. We will also follow this approach, and use the term motifs regardless of frequency.

For a *motif analysis* of a set of graphs, a set of possible motifs is selected a priori. These motifs must be unique, therefore pairwise non-isomorphic. To analyze a graph $G = (V, E)$, the occurrences of all motifs in $G$ are counted. An *occurrence* of a motif $M$ in $G$ is a set $X$ of nodes of $G$ such that the connected subgraph of $G$ induced by $X$ is isomorphic to $M$. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two directed graphs such that

$|V_1| \leqslant |V_2|$, and let $V_2' \subseteq V_2$ such that $|V_2'| = |V_1|$. Then $G_1$ is the *subgraph* of $G_2$ *induced by* $V_2'$ if there is a bijection $\varphi : V_1 \to V_2'$ such that for all $x, y \in V_1$, it is $(x, y) \in E_1$ if and only if $(\varphi(x), \varphi(y)) \in E_2$. Since no two motifs are isomorphic, each set of nodes of $G$ can be the occurrence of at most one motif. Even some motifs appear to be symmetrical, such as (1), (2), (6), (8), (9), (10), and (13) in Figure 3.6, the underlying node set is counted as exactly *one* occurrence. The number of occurrences of all motifs in a graph $G$, normalized to a sum of 1, is called the respective *motif signature*.

*Motif analysis* connects the structure and components of a graph to its semantic properties, depending on the specific application and context. Evaluation of motifs can be of quantitative or qualitative nature. *Quantitative* analysis focuses on the statistical relations between the motif frequencies and the semantic properties. In contrast, *qualitative* analysis tries to interpret the motifs, so that motifs of unusual frequency may reveal a deeper understanding of the underlying semantic properties.

Motifs capture local structure and are thus, in a sense, on an intermediate level between measures on single nodes and edges on one hand and global measures on the other hand.

Motif analysis has first been investigated in computational biology [76] and has since been applied to a variety of network types in biology and biochemistry [74]. The underlying insight is that biological and biochemical dynamics are statistically related to the occurrence of small functional blocks, which have specific structures. This insight is well captured by motif signatures, and in fact, many computational studies reveal significant relations. Due to this success, it did not take long time until this technique has been applied to graphs from other domains. For example, [60, 61] compare graphs from biology, electrical engineering, natural language and computer science and find

that the motif signatures from different domains are so different that they may serve as a “fingerprint” of the respective domain.

Krumov et al. [51] use motif analysis on co-authorship networks to find relations between particular motifs and citation frequency. They reveal one particular motif that implies high average citation frequency, and provide explanations based on social processes that are covered by the graph.

Tran et al. [80] explored differences in directed and undirected networks of various disciplines, including ecology, biology and social science. They conclude that motifs in undirected graphs are very similar. However, motif analysis of directed graphs was able to distinguish graphs from different fields. Furthermore, larger motifs captured more information about individual differences than small motifs.

Quite recently, motif analysis has been applied to text processing. In the research of Biemann et al. [18], human-written texts and artificial texts with quite similar characteristics were compared by means of the motif signatures of certain induced graphs. For several natural languages, the motif signatures were so different that they alone were sufficient to distinguish the human-written from the artificial texts. In an extension of their work, they identify significant differences in motif signatures that are restricted to co-occurrence of verbs, predicates, and other word classes, and present results on peer-to-peer streaming, co-authorship, and mailing networks [19].

Mesgar and Strube [58] applied motif analysis to Wall Street Journal news articles. These texts were represented as a combined entity and discourse relation graph. They identified several motifs that are highly correlated with manual readability ratings, and are able to transfer this knowledge to improve the task of machine translation [24].

# 4 Assessing Text Quality with Motif Analysis

## 4.1 Introduction

Recent work has shown that *motif analysis* is quite promising for natural language processing [18, 58]. Roughly speaking, the occurrences of small subgraphs like those in Figure 3.6 are counted, and relations between this *motif signature* and the semantics of the network are analyzed. In the following chapter, we will demonstrate that motif analysis can help in the assessment of text quality. Our computational study is based on the German Wikipedia. The label "featured" indicates articles of particularly high quality. The length (number of words) of an article is a comparatively good predictor for this label [22]. We show that a well-designed combination of this criterion and motif statistics yields a significant improvement. We also find that a deeper look into the most relevant motifs may improve our understanding of quality.

This chapter is organized as follows. First we briefly sketch our contribution in Section 4.2. Then, in Section 4.3, we discuss the state of the art. We illustrate the composition of our corpus in Section 4.4 and details of our approach in Section 4.5. Quantitative and qualitative results are presented in Section 4.6 and 4.7, respectively. In Section 4.8, we show the Motif Analysis Toolkit - a Java framework that unifies all processing steps of this experiment. We use this software to apply this workflow to other scenarios, and briefly discuss the results. Finally, we summarize our work in Section 4.9.

The results of this research have been published at the TextGraphs Workshop as part of the North American Chapter of the Association for Computational Linguistics 2016 in San Diego [12].

## 4.2 Our Contribution

In this chapter, we address one main research question of the thesis: Can we use graph motifs to assess text quality? In more detail, we discuss three specific questions:

1. *Quantitative*: Does motif analysis as a stand-alone tool help us assess the quality of text documents statistically?
2. *Quantitative*: Does it help us in conjunction with other quality measures?
3. *Qualitative*: Does it help us understand the nature of quality any better?

The German Wikipedia is our basis. The Wikipedia allows the community to assign the label "featured" to an article via an extensive communication and revision process, based on a collection of stylistic and content-based quality criteria. We use the distinction featured / non-featured as a (binary) quality criterion. Our corpus comprises all featured articles and a purely random selection of non-featured articles such that 7% of all articles in our corpus are featured.

In summary, we address the specific research questions of this chapter in the following form: Does motif analysis help us – alone or in conjunction with another criterion – to distinguish between featured and non-featured articles, and if so, does it yield a deeper understanding of the nature of featured articles?

The bare length of an article is a surprisingly good predictor for whether or not an article is featured. So, for the second research question, we will combine the article length with our motif analysis.

The revision history of a Wikipedia article allows us to analyze the development of the induced graph and its motif signature over time. So, for each article we analyzed a series of "snapshots" and the temporal tendency of the motif signature.

## 4.3 Related Work

Defining and measuring the quality of a document in a formalized way is an intrinsically difficult task. Various mathematical measures have been proposed for individual aspects of quality, like correct grammar [79] or spelling [28]. Another part of quality, information ordering, has been evaluated with rank correlation metrics [53]. Louis and Nenkova [55] investigated text quality at a much broader level that incorporates features of emotions and surprise. A detailed depiction of quality flaws and a model for article quality in Wikipedia was presented by Ferschke [37]. In contrast to our work, a large number of lexical, semantic and meta-data features were used for quality flaw detection, including named entities, readability assessment and links to other Wikipedia pages and external sites. We want to find general textual motifs without using specific Wikipedia characteristics and structures.

A graph based approach on quality assessment of school essays was presented by Antiqueira et al. [7]. This research used global statistical network features of text-induced graphs, including mean clustering coefficient and average shortest path length. The authors presented correlations between these metrics and manually annotated quality scores.

## 4.4 Data

Our corpus is a subset of the German Wikipedia. We utilize the quality label "featured" as an indicator for high quality articles. Therefore, we include all 2,338 featured articles

of a complete snapshot of the German Wikipedia from June 2015. The proportion of featured articles in this snapshot is 0.13%. Adding all non-featured articles would result in a dataset of very large proportions and with an extremely skewed distribution of the relevant "featured" label. This is a huge problem for most machine learning techniques. We deal with this problem by under-sampling the overrepresented class [34]. For this reason, we restricted the set of all non-featured articles to a purely random – and thus representative – sample of 33,295 articles, which increases the share of featured articles in our corpus to 7%.

For each article in our corpus, we selected 10 distinct article versions from the article's revision history. A new version of an article is created every time submitted revisions or additions to this article are approved by the community. For every featured article, we split all of its article versions into a set of versions with the featured label and a set of version without the featured label. On average, this divides the versions of a featured article into about 57% non-featured versions and 43% featured versions. For every featured article, we select five random versions from each part.

We split the versions of non-featured articles similarly into "early" and "late" parts of the revision history with the same proportions as the featured articles' featured and non-featured parts. This way, we pick five random versions from the earliest 57% versions of a non-featured article, and five random versions from the latest 43% versions.

For every article version, we create a graph according to our graph representation, search the graph for motifs and compute the corresponding motif signatures, as explained in the following Section 4.5.

## 4.5 Our Approach

We use a sentence level graph representation based on shared nouns. For every selected version of a selected Wikipedia article, we construct a graph $G = (V, E)$ with a set of nodes $V$ and a set of directed edges $E$ as follows: The nodes of this graph are the sentences of the article version. Two nodes are connected by an edge if and only if these two conditions are fulfilled:

1. There exists at least one noun token that appears in both corresponding sentences.
2. The two sentences are separated by at most two other sentences in the document.

Edges are directed and point from the sentence earlier in the text to the latter one. Figure 4.1 shows an example of this representation.

To each graph created from a Wikipedia article version, we apply a motif analysis. In our case, we search for subgraphs of three or four connected nodes. Furthermore, we only search for motifs of three or four directly consecutive sentences. With this constraint, we can only discover discourse connections of sentences that follow right after each

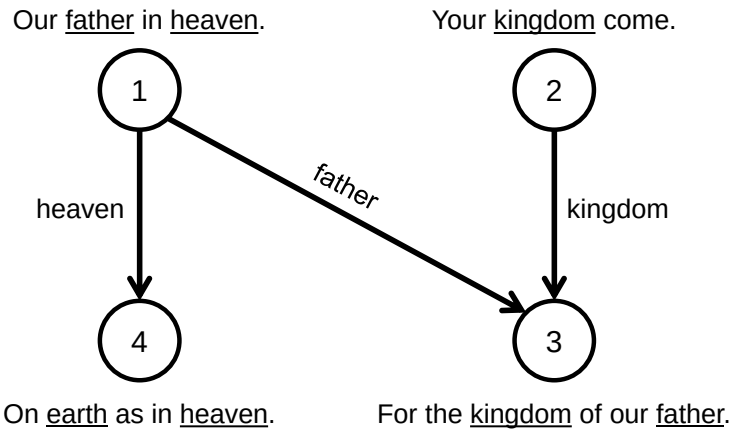



**Figure 4.1:** Exemplary graph representation with four consecutive sentences. Noun tokens are underlined. In this visualization, edges are labeled with the matching noun tokens of the connecting sentences.

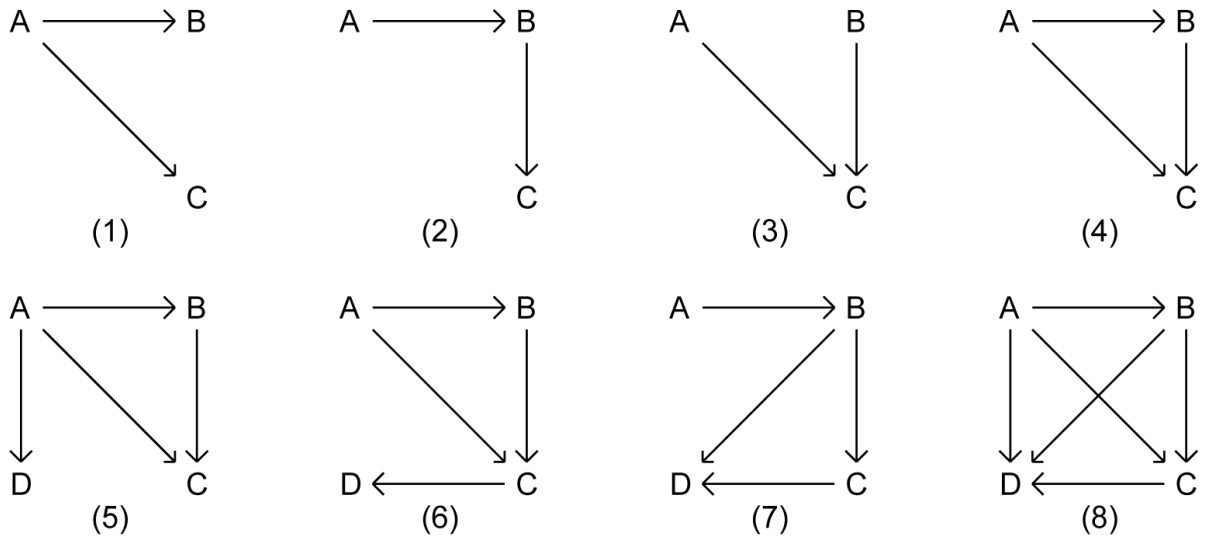



**Figure 4.2:** All possible directed motifs on three nodes, and four selected motifs on four nodes.

other. Because of their close proximity, we can be pretty sure that these sentences have a strong discourse connection and are likely to share a common topic to some extent.

The resulting motifs are quite rare. If we relaxed this constraint and searched for all connected subgraphs of three or four nodes, the number of occurrences of motifs increases significantly. However, we found that the motif analysis yields worse results.

The way our graphs are constructed limits the number of possible subgraphs considerably. The directions of the edges have to follow the order of appearance of the corresponding sentences, which rules out every form of loops. All four possible connected subgraphs on three nodes and four subgraphs (out of 42) on four nodes are shown in Figure 4.2.

The node order together with the adjacency condition allows for very efficient searching for these motifs with a sliding window. The occurrences of all three-node motifs and all four-node motifs are scaled to a sum of 1, respectively. The results build the motif signatures, as defined in Section 3.2.

We use the values of the motif signatures as 46 numeric features (four three-node motifs + 42 four-node motifs) for our machine learning experiments. In addition, we include the word count of the article version as an additional baseline feature according to Blumenstock [22], for comparison and combination. The experiments were performed with J48, a Java implementation of the C4.5 tree learning algorithm, included in the Weka machine learning toolkit [70, 87]. The tree structures allow us to interpret the model and analyze the most determining features. We use default parameters with the exception of "minNumObj", the minimum number of instances per leaf. The default value of this parameter is 2. We set it to 100 to reduce overfitting effect, and will present results for both configurations. The evaluation is performed with 10-fold cross validation over 10 experiment iterations.

## 4.6 Quantitative Results

Our corpus contains 7% featured article versions. Therefore, consistently predicting the majority class "non-featured" produces a lower bound baseline accuracy of 93%. The baseline we want to compare with is created by a J48 experiment with the word count feature only, which achieves 95% accuracy.

We evaluate the predictive power of our motifs with experiments that use only 3-node motifs, only 4-node motifs or both. The results for these experiments with default number of 2 instances per leaf are presented in Table 4.1. Experiments with all 3-node motifs without word count could not reach the baseline, but using all 4-node motifs without word count performed much better at 97.45% accuracy. This includes significant overfitting effects, as the corresponding tree model is very large and consists of over 3,000 nodes and leafs.

| Used Features | Accuracy | Tree size |
|---|---|---|
| Majority class baseline | 93.00 | |
| 3N + 4N | **97.48** | 3605 |
| 4N | 97.45 | 3598 |
| W (baseline) | 95.00 | 6 |
| 3N | 94.88 | 1417 |

**Table 4.1:** J48 results for article quality predictions with motifs alone, parameter minNumObj = 2, W = word count, 3N = all 3-node motifs, 4N = all 4-node motifs.

| Used Features | Accuracy | Tree size |
|---|---|---|
| Majority class baseline | 93.00 | |
| 3N + 4N | **95.08** | 217 |
| 4N | 95.04 | 188 |
| W (baseline) | 95.00 | 3 |
| 3N | 94.30 | 140 |

**Table 4.2:** J48 results for article quality predictions with motifs alone, parameter minNumObj = 100, W = word count, 3N = all 3-node motifs, 4N = all 4-node motifs.

Table 4.2 shows the results with 100 minimum instances per leaf (up from 2), which reduces the model size and the overfitting effects. Lower bound and baseline accuracy are the same as in the first setup, at 93% and 95%, respectively. In this setup, using only 3-node motifs yields an accuracy of 94.30%. 4-node motifs alone or in conjunction with 3-node motifs do not outperform the word count baseline considerably, either. We conclude that our motif analysis as a stand-alone tool did not lead to notable statistical improvements.

In our experiments, we also combined the baseline feature word count with motif features. The results for these combinations are shown in Table 4.3 with default number of 2 instances per leaf and Table 4.4 with 100 instances per leaf. At default setup, a com-

| Used Features | Accuracy | Tree size |
|---|---|---|
| Majority class baseline | 93.00 | |
| W + 3N + 4N | **97.80** | 3179 |
| W + 4N | 97.79 | 3254 |
| W + 3N | 95.67 | 847 |
| W (baseline) | 95.00 | 6 |

**Table 4.3:** J48 results for article quality predictions with feature combinations, parameter minNumObj = 2, W = word count, 3N = all 3-node motifs, 4N = all 4-node motifs.

| Used Features | Accuracy | Tree size |
|---|---|---|
| Majority class baseline | 93.00 | |
| W + 3N + 4N | **96.00** | 179 |
| W + 4N | 95.77 | 200 |
| W + 3N | 95.42 | 59 |
| W (baseline) | 95.00 | 3 |

**Table 4.4:** J48 results for article quality predictions with feature combinations, parameter minNumObj = 100, W = word count, 3N = all 3-node motifs, 4N = all 4-node motifs.

bination of word count with 3-node and 4-node motifs shows excellent performance at 97.80% accuracy. Again, very large decision tree models were created, which indicates overfitting. Limiting the tree size reduces the decision tree to very moderate size and still results in 96% accuracy. Motifs together with word count close the gap from the baseline's 5% miscategorized examples to 4%.

We confirmed this improvement and its statistical significance on reduced subsets of our data, and also in a balanced setting. We created the reduced subsets by a purely random selection of 10% featured and non-featured article versions. Compared to the full dataset, these subsets have reduced size, but the same ratio of featured vs. non-featured instances. To create balanced subsets, we combined all featured article versions and an

| Used Features | Mean acc. | Standard dev. |
|---|---|---|
| W + 3N + 4N | 95.116 | 0.073 |
| W + 4N | 95.093 | 0.071 |
| W + 3N | 95.079 | 0.085 |
| W (baseline) | 94.741 | 0.096 |

**Table 4.5:** Mean accuracy and standard deviation for 20 reduced datasets, J48 parameter minNumObj = 100, W = word count, 3N = all 3-node motifs, 4N = all 4-node motifs

| Used Features | Mean acc. | Standard dev. |
|---|---|---|
| W + 3N + 4N | 94.963 | 0.111 |
| W + 4N | 94.941 | 0.124 |
| W + 3N | 94.852 | 0.107 |
| W (baseline) | 94.440 | 0.109 |

**Table 4.6:** Mean accuracy and standard deviation for 20 balanced datasets, J48 parameter minNumObj = 100, W = word count, 3N = all 3-node motifs, 4N = all 4-node motifs

equal amount of randomly selected, non-featured article versions. We constructed 20 subsets for both settings to measure variance and calculate statistical significance. See Table 4.5 for the results of the reduced subsets, and Table 4.6 for the results of the balanced scenario.

Due to the reduced amount of data, the mean accuracy was lower compared to previous experiments with same features and parameters, but the order remained constant. The accuracy is very stable with respect to data composition, with standard deviations between only 0.073 and 0.124. The mean accuracy of every combination setup surpassed the baseline by at least three standard deviations. Computation of p-values confirmed that all results are statistically highly significant at $p < 0.001$.

We obtain the motifs with highest impact on quality with two different methods.

| Motif | Accuracy |
|---|---|
| (4) | 95.1713845 |
| (8) | 95.1328903 |
| (3) | 95.0983271 |
| (1) | 95.080967 |
| (7) | 94.9560177 |
| (2) | 94.9386417 |

**Table 4.7:** Accuracy of J48 experiments using only single motifs and word count as features.

To determine the most effective motifs in our machine learning setup, we performed additional experiments with single motifs in combination with word count. Table 4.7 displays the results with highest accuracy, which indicates a connection between these motifs and quality. The motif labels correspond to Figure 4.2.

As a second approach, we directly evaluated the correlation of the motif signatures to the quality label of the corresponding text. Since our variable for quality is binary, we use the point biserial correlation coefficient. The value distribution of our motif signature entries imposes potential problems for this computation. An example of this distribution is shown in Figure 4.3. We see large proportions of the extreme values 0 and 1 in our motif signature distributions. A value of 1 in the signature can only happen if the respective motif was the only motif to be found, which only happens in very small texts. A 0 entry also hints to small texts, as large texts tend to contain at least a fraction of every motif type. Small articles in Wikipedia are rarely featured, so both extreme values of the distribution largely correspond to non-featured articles, which distorts the point biserial correlation coefficient. To eliminate the effects of article length in the correlations, we create sub-collections of featured and non-featured article versions with similar amounts of motifs, measured by mean average values and standard deviation.

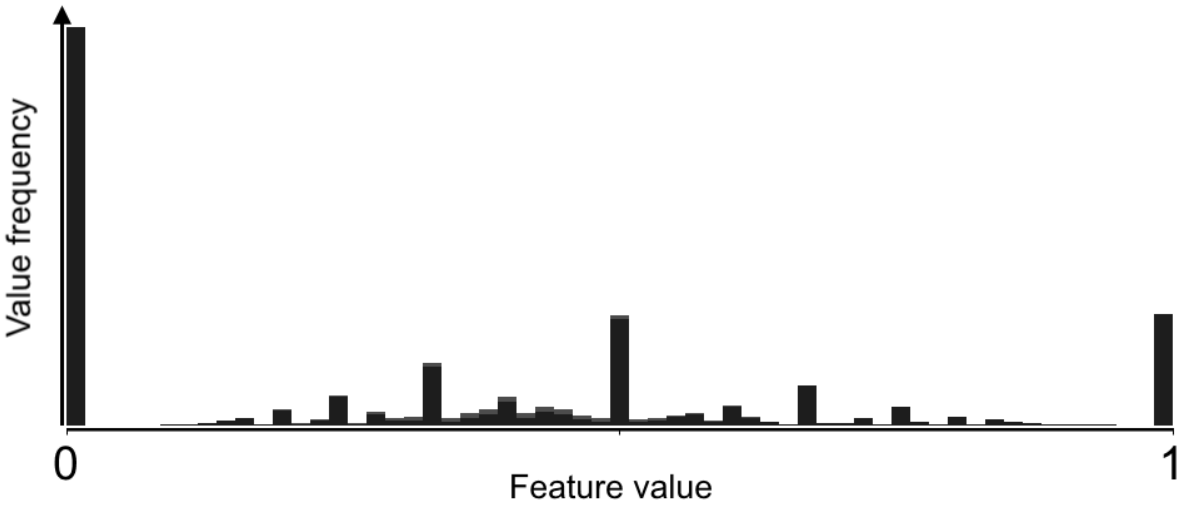



**Figure 4.3:** Frequency distribution of feature values for motif (4) (see Figure 4.2) over all motif signatures.

| Motif | Correlation |
|---|---|
| (1) | 0.2015307459 |
| (3) | 0.1847741318 |
| (2) | 0.1803760107 |
| (8) | -0.2691052665 |
| (4) | -0.27433768 |

**Table 4.8:** All motifs with correlation coefficient of absolute value $> 0.15$.

We found two motifs with a distinctively negative correlation to the featured quality label, and three motifs with distinctively positive correlation coefficients. See Table 4.8 for these results. All these motifs have also been identified by our machine learning approach with single motifs (see Table 4.7), which confirms our findings. We concentrated our qualitative interpretation on the motifs with most dominant results, which are motifs (1), (3), (4) and (8) (see Figure 4.2).

## 4.7 Qualitative Results

The two motifs that are most positively correlated with the featured quality label in both methods are motifs (1) and (3) (see Figure 4.2). The huge amount of data made an exhaustive investigation of all motif occurrences infeasible. Therefore, we examined random examples of these motifs in our data set.

Many examples of motif (1) showed this sentence structure: The first sentence introduces two connected entities; the second sentence offers details about the first entity; the third sentence explains the other entity. One example from the German Wikipedia, and a translation into English:

(a) Von den drei Hartstrahlen der Rückenflosse wurde der erste zur Angel (Illicium) mit anhängendem Köder (Esca) umgebildet.
(b) Das Illicium hat oft eine Streifenzeichnung.
(c) Die Esca hat bei den einzelnen Anglerfischarten eine unterschiedliche Form und ist ein wichtiges Unterscheidungsmerkmal zwischen den Arten.

(a) Of the three hard rays of the dorsal fin, the first was rebuilt to Angel (Illicium) with attached bait (Esca).
(b) The Illicium often has a stripe drawing.
(c) The Esca has a different shape in the individual anglerfish species and is an important distinguishing feature between the species.

Analyzing motif (3) revealed a similar, but decisively different pattern: The first sentence introduces one entity, the second sentence introduces a second one. The last sentence combines the two mentions and draws a connection.

(a) Die Haupthenne bleibt mit dem Hahn oft über mehrere Jahre zusammen.

(b) Bei den Nebenhennen handelt es sich meistens um recht junge Weibchen.

(c) Der Hahn paart sich zunächst mit der Haupthenne, sodann mit den Nebenhennen.

(a) The main-hen often stays with the rooster for several years.

(b) The side-hens are mostly young females.

(c) The rooster first pairs with the main-hen, then with the side-hens.

These two ways to introduce and connect entities are an indication of good writing style. The reader can see explanations for the two mentioned entities and their connection in direct vicinity. This makes it easy to understand and to follow the argument structure. Motifs (4) and (8) are highly negatively correlated with the Wikipedia article quality. They share a very similar structure: Motif (4) is a maximally connected 3-node subgraph, motif (8) is a maximally connected 4-node subgraph. Many text examples of these motifs share a pattern of repetition. One noun is used three or four times in a row in very close proximity.

(a) Die Nahrungssuche erfolgt in der Regel einzeln, seltener als Paar.

(b) Bei Beobachtungen waren die Tiere in 92 % aller Fälle allein auf Nahrungssuche.

(c) Beutejagd und Nahrungssuche bestehen vor allem darin, dass die Füchse ihre Beute zumeist zwischen und unter Steinen suchen und gelegentlich auch graben.

(a) Foraging is usually done individually, rarely as a couple.

(b) In observations, the animals approached foraging alone in 92% of all cases.

(c) Prey hunting and foraging consist mainly in the foxes searching and sometimes digging for their prey between and under stones.

Repetition is a strong stylistic device that can enhance learning effects, but it can also make the text tedious and reduce the attention of a reader [44, 78]. Too much repetition

is a sign of bad writing style and is certainly avoided in good articles, as our findings demonstrate.

## 4.8 Motif Analysis Toolkit

The experiments on Wikipedia data followed two objectives. First and foremost, they directly cover one important research question of this thesis, and were carefully chosen to discover potential connections between text induced graph motifs and text quality. Second, the workflow of the experiments can be seen as a prototypical application of motif analysis on text without additional meta data or external knowledge bases. Therefore, we designed a framework to combine all necessary processing steps in an easily usable toolkit.

The Motif Analysis Toolkit is an open source Java software that follows two important design requirements: Usability and expandability. The full processing workflow from raw source text to graphs, motifs and machine learning results with all implemented functions should be possible without any programming expertise. Although programming skills are needed to expand the toolkit, the required changes should be as minimal as possible.

To enable easy access to all features, we added an intuitive graphical user interface to the toolkit [45]. For expandability, every processing step is implemented with generic and abstract modules. This way, changing the text parser or the graph representation can be realized by adding one additional java class.

In the following sections, we will demonstrate the main features of this toolkit, and present additional experiments that were conducted with its help. The software is licensed under the GNU General Public License v3.0.[1]

[1] https://bitbucket.org/Arnoldex/aiphes_motifanalysis

### 4.8.1 A Quick Tour

The graphical user interface of the Motif Analysis Toolkit consists of three main views for the most important processing and analysis steps. The main view (see Figure 4.4) is the starting point of the full workflow. The user chooses paths for input and output files, and additional parameters, like the language of the input texts or the used graph representation. The standard input format is plain text. The start button in this view launches the motif analysis pipeline - all input texts are transformed into a graph representation according the the chosen parameter, and all graph are searched for graph motifs of size three and four. All results, including the motif signatures in Weka compatible .arff format and graph .gml files, are saved in the specified output folder.

The second view (seeFigure 4.5) contains options for a quick machine learning evaluation. The user can specify an .arff file that was generated in the previous step, and run a selected machine learning classifier. A summary of the results is shown in the console section of the GUI. If further modification, preprocessing steps or unsupported classifiers are needed, Weka can be started directly from the view.

The last view (see Figure 4.6) can be used to visualize selected graphs and get a general idea of its structure. The toolkit uses the external functionality of Gephi, a free graph software under the GNU General Public Licence [1]. The user can choose a .gml file that was generated in the first step of the computation pipeline, and get a quick visualization generated by Gephi. Figure 4.7 shows an example visualization. In the example, every sentence is modeled as one node, similar to the experiment in this chapter. The nodes are numbered, and represent the sentence order. Edges between two nodes indicate same nouns, and the edge labels reveal the matching words.

[1] https://gephi.org/

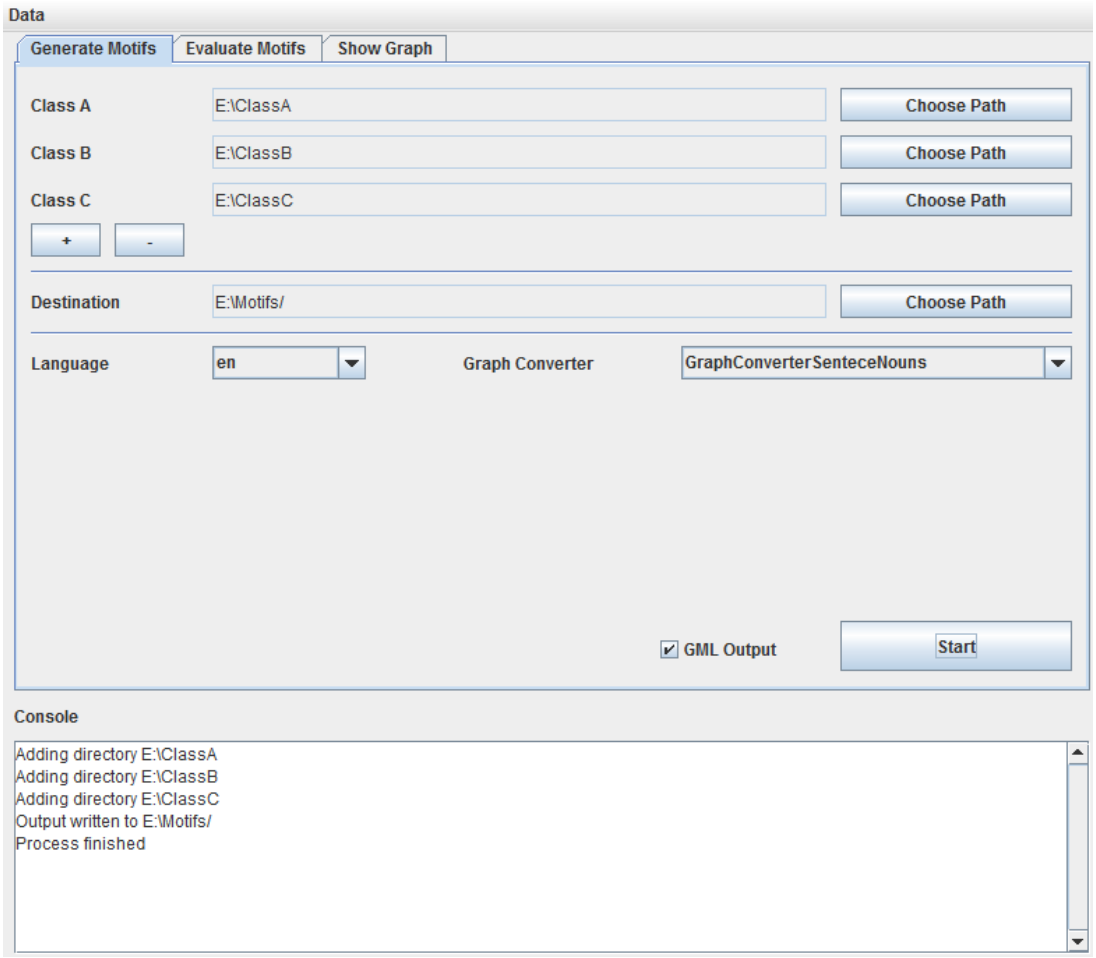



**Figure 4.4:** The main view of the Motif Analysis Toolkit. The user can define multiple input paths for data of different classes, and choose several parameters. Clicking the start button transforms all input documents into graphs, searches these graphs for motifs, and saves all results in the output folder.

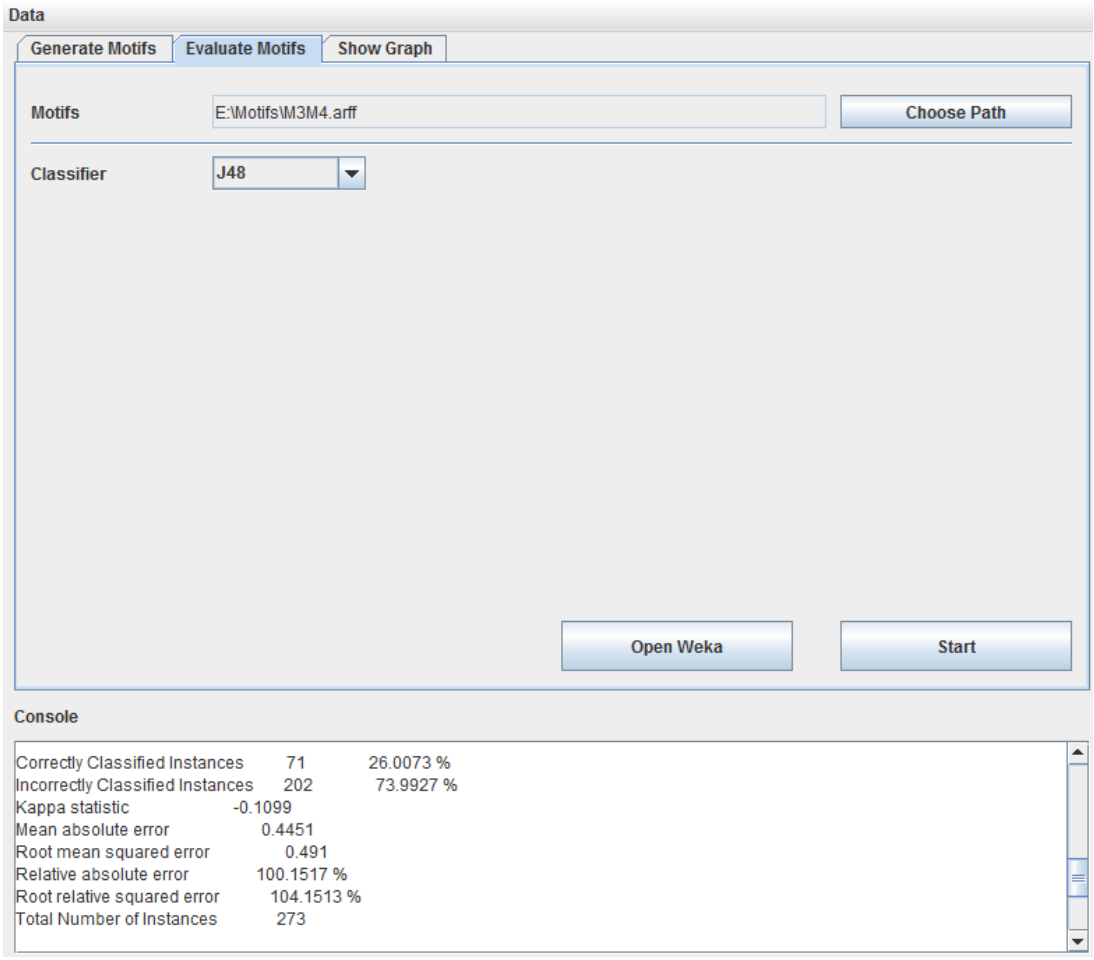



**Figure 4.5:** The Weka view of the Motif Analysis Toolkit. The user can analyze a certain set of features (motifs of different sizes, baseline features like word length) in a simple classification setup.

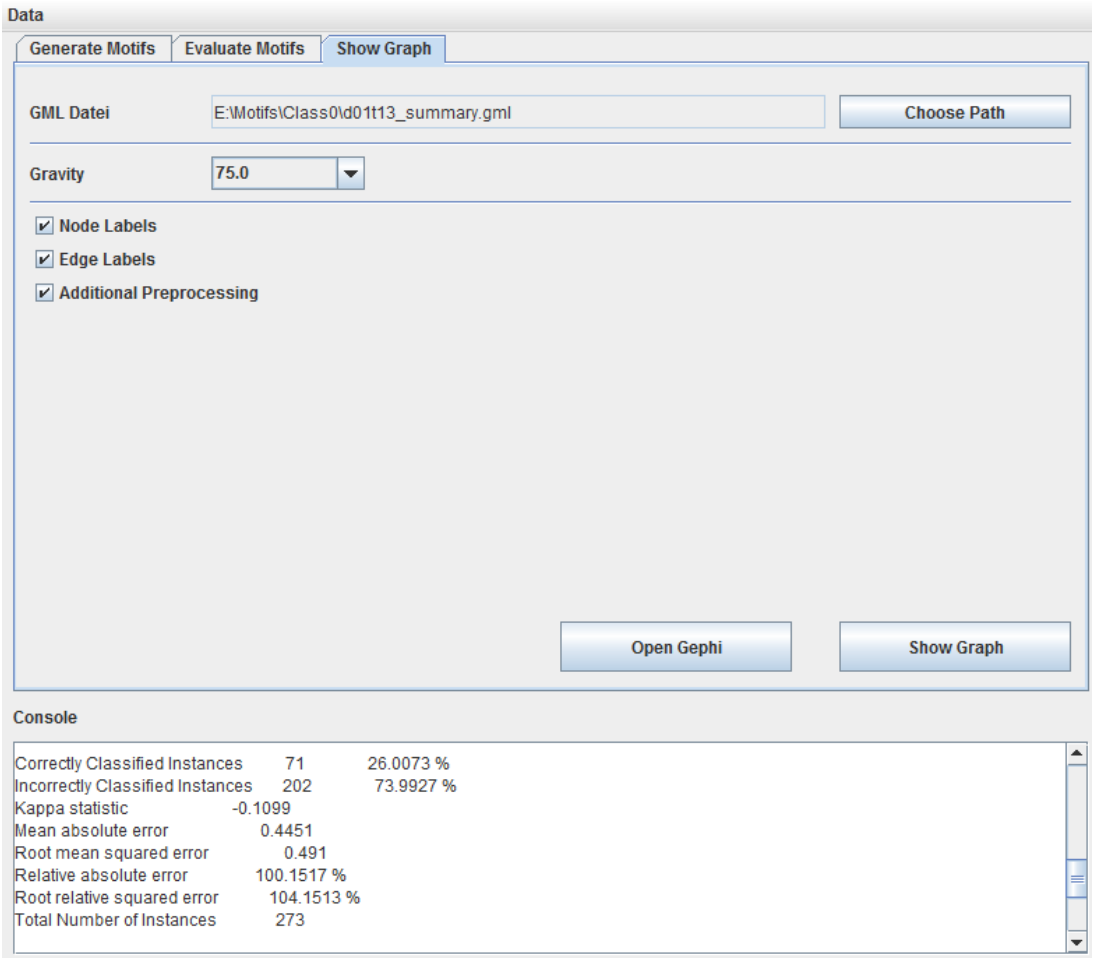



**Figure 4.6:** The Gephi view of the Motif Analysis Toolkit. The user can look at visualizations of selected graphs.

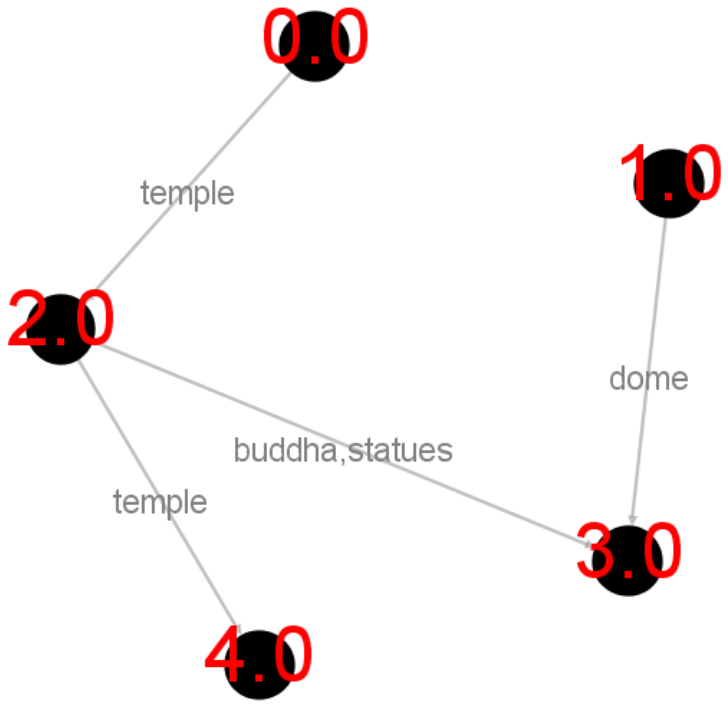



**Figure 4.7:** A sample visualization of a small graph, created with the Motif Analysis Toolkit.

### 4.8.2 Applications on Other Data Sets

With the assistance of Motif Analysis Toolkit, we tested three additional applications of the described motif analysis workflow. Each setup uses the same graph representation: The nodes of this graph are the sentences of the input text. Two nodes are connected by an edge if and only if these two conditions are fulfilled:

1. There exists at least one noun token that appears in both corresponding sentences.
2. The two sentences are separated by at most two other sentences in the document.

All classifications use the J48 tree learning algorithm with 10 runs of 10 fold cross validation. The following sections briefly describe each additional experimental setup and its results.

### Shuffled vs. Unshuffled Summaries

The data set of this experiment consisted of a 91 summary subset from the hMDS corpus [93]. The summaries in this corpus were extracted from the first paragraphs of English Wikipedia articles. We hypothesize that motifs can distinguish between summaries with a correct sentence order from summaries with randomly shuffled sentences. We create the shuffled data set by randomizing the sentence order of the 91 original summaries. This way, we get a balanced data set. A random classifier is expected to achieve 50% accuracy.

Using motifs of size three and four as features, the classifier yielded 54.3% accuracy, slightly above the baseline. Given the small data size of 91 input documents, this improvement is not statistically significant. One explanation of these results can be found in the size of each individual summary. Since each summary is a short paragraph, they only contain between 5 and 20 sentences, and 191 words on average. Therefore, each graph consists of very few nodes, and in consequence, very few motifs. This can be seen in very sparse motif signatures in the data set. We assume that motif analysis, as a statistical method, is ill-suited to very short input documents, at least in this form of application.

### Positive vs. Negative Movie Reviews

The assumption that short texts do not work well with motif analysis was strengthened by an experiment of movie reviews. We use the movie review dataset by Maas et. al [56], with 25,000 highly polar movie reviews extracted from the Internet Movie Database (IMDb)[1]. The reviews are labeled to be positive or negative, with an equal dis-

[1] www.imdb.com

tribution between these two labels. We tested motifs as features to predict the positive or negative sentiment label of these movie reviews.

The results were very similar to the results of the previous experiment. With the motifs, 54% of all movie reviews were classified correctly, which is only slightly better than the random baseline of 50%. Looking at the input data and the results, we see that the reviews are even shorter that the Wikipedia summaries, with a mean of 178 words per document. The much larger amount of documents confirms the previous assumptions. The analysis might work better with a different, more fine-grained graph representation, that goes deeper than sentence level.

### Real vs. Fake Science Articles

This third example for motif analysis application uses scientific articles, both real articles and automatically generated fake articles. Mathgen[1] creates randomly generated research papers that look like maths articles. SCIgen[2] is a similar program for random computer science papers. Another webpage creates nonsense articles in the field of Postmodernism[3]. Without reading the text, the results of all three generators looks reasonable. Automatically generated figures and tables are included, also (random) citations and a bibliography. We compare these hoax articles with real articles from a subset of the KDD Cup 2003 data set [42]. This corpus contains extracted arXiv papers mainly from the field of physics. Our subset consists of 391 real papers from this data set, and we used the three hoax paper generators to create roughly 10,000 random papers each.

1 http://thatsmathematics.com/mathgen/
2 https://pdos.csail.mit.edu/archive/scigen/
3 http://www.elsewhere.org/journal/pomo/

| **Mathematics** | **Postmodernism** | **CS** | classified as |
|---|---|---|---|
| 10000 | 0 | 1 | **Mathematics** |
| 2 | 10032 | 65 | **Postmodernism** |
| 0 | 8 | 10091 | **CS** |

**Table 4.9:** Confusion matrix of hoax paper classification experiment.

As a first classification experiment, we use only the 30,100 generated papers. The target prediction in this setup is the topic of the input paper - mathematics, computer science, or postmodernism. In other words, we want to predict the used generator. This worked extremely well, as proven by the confusion matrix in Table 4.9. The overall accuracy is 99.75%, with a random baseline of 33.55%.

In additional setups, we focus on the distinction between real and generated papers. In setup A, we use three different classes for the three different paper generators. In setup B, we only consider the classes “generated” versus “real”, and use one combined class for all generated documents, regardless of the used generator. Finally, we exclude postmodernism papers in setup C, with the intuition, that the real physics papers are much closer to the other two topics, therefore the classification might be harder. We again combine the hoax papers from the other two generators to a shared class.

The results are presented in Table 4.10. In all three setups, the classification works extremely well, with perfect or almost perfect accuracy. Looking at the code and algorithms for the paper generators, we quickly see an explanation for these results. All generators use fixed lists for vocabulary, generation rules and overall structure patterns. As a consequence, some motifs can appear quite often in one generator, but is simply impossible in another one. Therefore, the motif signatures of such automatically generated documents are easily distinguishable, and the classification of different generators is extremely easy. Real papers do not follow these strict rule sets, and have much more

| Setup | Accuracy |
|---|---|
| A | 99.20 |
| B | 99.40 |
| C | 100.00 |

**Table 4.10:** Accuracy of three hoax paper classification setups. Setup A uses separate classes for each of the three generators, setup B combines all generated papers into one class. Setup C does the same, but only includes documents from the mathematics and computer science generators.

organic motif signatures. Thus, the classification of real and generated articles is also trivial. Another observation is the increase in accuracy when excluding the papers created with the postmodernism generator. Looking at the motif signatures, these excluded papers showed much higher variation compared to texts from the other two generators. Their generator creates "more random" output, with has a higher chance of being similar to a real paper. The rule sets of the other two generators is more limited, and the motif resulting signatures are very similar to other ones from the same generator. The used tree learner was able to detect these peculiar motif signatures with ease.

## 4.9 Conclusion and Outlook

We have seen that motif analysis can improve the assessment – and our understanding – of the quality of a document. For that, we explored one particular setting and presented the results in this chapter. We formulated the following research questions:

1. *Quantitative*: Does motif analysis as a stand-alone tool help us assess the quality of text documents statistically?
2. *Quantitative*: Does it help us in conjunction with other quality measures?
3. *Qualitative*: Does it help us understand the nature of quality any better?

In our corpus, only 7% of all articles are featured. Hence, categorizing *all* articles as non-featured gives quite a high base line. If the threshold is well chosen, the word count of an article miscategorizes 5%.

Our motif analysis alone is not better than that, so the answer to the first research question is not strictly positive. However, we showed that our combination of both criteria reduces the share of miscategorized articles from 5% down to 4%.

For our third research question, we identified a subset of motifs with high positive or negative correlation to the featured label. Two motifs occur outstandingly frequently in the featured articles, and two other motifs occur outstandingly frequently in the non-featured articles. All four motifs are indeed indicators of text quality as desired: the two former ones are frequently induced by two concepts of good writing style, whereas the two latter ones are frequently induced by two cases of repetitive style.

We combined the general motif analysis work flow that has been developed for this experiment in an easily usable and expandable open source Motif Analysis Toolkit. Using this software, we conducted additional experiments on three different data sets. In these applications, motifs were not able to distinguish randomly shuffled summaries from the original documents, or positive versus negative movie reviews. In both cases, the documents were quite small, which is a problem for statistical methods. In another application, motifs easily distinguished real scientific articles from automatically generated papers. The generators followed strict generation rule sets. As a consequence, the motif signatures from such generated texts did not vary much, and could be told apart from real data with ease.

Motivated by these results, we will now investigate the effects of motif analysis on quality on a very different level.

# 5 User Interaction Motifs and Community Performance

## 5.1 Introduction

In Chapter 4, we have demonstrated the benefits of motifs on surface level text graphs of Wikipedia articles. Now, we test the feasibility of motif analysis on very different, yet still Wikipedia based graphs. This study still addresses the first main research question of this thesis: Can we use graph motifs to assess text quality?

In collaborative online writing communities, like Wikipedia, users are often categorized by a selection of informal roles. These roles describe prototypical behavior and contribution patterns. For instance, a certain group of users might focus on adding new content to articles, or prefer small surface-level editing like typo correction. Although previous work examined the effects of informal roles within collaborative communities, so far, the effects of interaction between these user groups has been neglected. Since interaction between users is the main benefit and unique feature of collaborative writing, we want to approach this gap with motif analysis.

In this work, our graphs are based on the informal roles and user interactions in several online writing communities, namely Wikipedia and Wikia. We measure the effect of reoccurring user interaction patterns - our motifs - on the overall community success and quality. Our results reveal that certain interaction patterns have a measurable effect on community success that cannot be found when only looking at the user groups and

roles in isolation. In particular, we discovered that cooperation of users that focus on content quality over quantity has a constant positive effect.

We will first describe our contribution in Section 5.2, and discuss connected research in Section 5.3. We introduce our main data source, the collaborative online platform Wikia, in Section 5.4. In the following Section 5.5, we describe our methods, and present step wise results. We combine and discuss our findings in Section 5.6, and close the Chapter with a summarizing conclusion in Section 5.7.

The results of this research have been published at the Wiki Workshop as part of the World Wide Web Conference 2017 in Perth [13].

## 5.2 Our Contribution

The importance of collaboration and user interaction in social networks and online communities is well known [54, 83, 86]. Since online writing communities, such as Wikipedia, are important public resources, they have attracted increased research on these topics. Some studies emphasize that successful collaborative processes require a number of experts [63, 68], whereas other results show that the most important factor of community success is the potential of many little contributions, even without particular background knowledge [49, 85].

Another stream of literature emphasizes that the benefits of crowd collaboration needs a high degree of coordination [8]. In particular, current research stresses the importance of organically emerging, implicit coordination [10]. In the context of collaborative writing communities, the typical edit behavior of contributors has been modeled in the form of informal roles to capture this implicit coordination [10, 54]. For instance, contributors with a tendency of inserting useless or disruptive content could be labeled

with the informal role “vandal”, whereas users that focus on correcting surface level mistakes, like typos and wrong grammar, may be considered “copy editors”.

Research on informal roles offers interesting insights about contributors and *what* they do, but nothing about *whom* they interact with, or in which way this interactions happens. However, a large number of studies cover the significance of collaboration and interaction in online communities [27, 52, 82]. This offers great opportunities to study informal roles, and especially the interactions and connections between them.

We hypothesize that the overall success and quality of online communities does not only depend on the performance of individual users or user groups, but to a greater degree on the interaction between contributors. We assume that interaction between very diverse groups with complementary abilities can have particularly beneficial effects on the collaborative results. For example, interaction between content creators and copy editors could be more desirable than only cooperation between one type of users.

To test our assumptions and measure interaction, we combine fine-granular analysis of contributor edit behavior and resulting implicit role allocation with motif analysis. Using motifs to capture typical user interaction patterns, we can not only quantify interaction in online communities, but also apply qualitative evaluation to describe and interpret the interaction and interacting contributors. We assess the importance of interaction in the knowledge production process and measure the effect of informal roles and user interaction on the overall quality of the outcome. For our main data set and investigation target, we decided to use the fan-based for-profit community Wikia.

Compared to Wikipedia (launched in 2001, approx. 2 million active users), Wikia[1] is a more restrictive community (launched in 2006, approx. 13,500 active users) with a commercial background and additional editing limitations.

[1] http://http://www.wikia.com/fandom

Our findings reveal significant differences in the collaborative behavior of different writing communities. We show that the open editing policy of Wikipedia incorporates a significant administrative overhead to prevent mischievous behavior, but also ensures a balanced community of contributors and sustainable collaborative structures. Using motif analysis, we detect important interaction patterns (our motifs) which characterize but also distinguish the collaborative work across different writing communities in Wikia. Our analysis indicates that the combination of contributors, their informal roles and their interaction in terms of graph motifs provides a consistent picture of community performance.

## 5.3 Related Work

Previous work on collaboration in online and social networks has extensively analyzed the interaction between contributors based on graph structures [83, 86]. Most of these looked into quantitative properties of co-author networks or subgraphs. For example, Sachan et al. [72] analyze social graphs on Twitter and in email conversations to discover smaller communities of contributors with shared interests. Brandes et al. [27] define co-author networks to visualize differences in the behavior of contributors and to reveal polarizing articles in Wikipedia. Their networks are based on positive and negative interaction of Wikipedia contributors in the form of delete and undelete actions. These approaches have two limitations. First, they largely ignore the temporal dimension. A static analysis of graph structures, however, can only reveal limited insight, as online communities and particularly social networks tend to evolve dynamically [47]. Second, as they are typically based on (social) links between contributors (such as followers, likes etc.), they do not take into account the informal roles of contributors.

The latter might however reveal important information about the implicit coordination inside the network.

Jurgens and Lu [47] address these concerns by integrating formal roles (e.g. admin, bot) and the temporal sequences of edits into their analysis of Wikipedia. With this approach they are able to identify four types of contributors' behavior with increasing or decreasing frequency over the course of time in Wikipedia's history. However, both their model of edit types as well as their model of contributor roles are pretty course-grained and capture rather high-level properties of the collaborative process.

Another stream of literature has analyzed informal (or social) roles in online communities. As opposed to formal roles [9], informal roles are not awarded by an authority, but they emerge organically. For example, the posting behavior of contributors on reddit has been used to identify roles such as the "answer-person" [29]. Welser et al. [84] describe four social roles played by Wikipedia contributors, based on a small-scale manual analysis of edit patterns and a larger-scale analysis of edit locations. They find that new contributors would quickly adapt to fit into one of those roles and that their notion of social roles implicitly models the "social" network of contributors, i.e. their interaction on Wikipedia talk pages.

In our approach, we adopted a slightly different notion of informal roles, based on contributors' edit history. It involves a fine-granular classification of Wikipedia edit types such as spelling corrections, content deletion or insertion [32]. This method has first been suggested and tested for the online community Wikipedia by Liu and Ram [54] and improved by subsequent work. For instance, Yang et al. [90, 91] present a method for automatic multi-label classification of Wikipedia edits into 24 types, based on a manually annotated sample. They identified eight roles based on editing behavior, involving a manual evaluation. The training data for edit categories used by Yang et al.

[91] is rather small, and the performance of their automatic edit classification algorithm is lower as compared to the revision-based classification approach presented by Arazy et al. [10] and used in this work.

With respect to the analysis of co-author networks, our work builds upon Brandes et al. [27]. However, in contrast to Brandes et al., we use informal roles of contributors to create more generalized networks, which enables us to search for universal interaction motifs. The exploitation of network motifs for analyzing collaboration in Wikipedia has previously been proposed by Jurgens and Lu [47] and Wu et al. [89]. The latter used their analysis of motifs to predict article quality. The approach proposed in this work is different from previous work on motif analysis in online collaboration in that we measure the impact of recurring motifs based on informal roles for entire communities rather than single articles.

## 5.4 Online Collaboration in Wikia

Wikis offer a convenient resource to study collaborative writing behavior as they have low entry barriers for new contributors, but at the same time they offer a reasonable administrative structure which allows to record and reverse any editing action. In the present study, we analyzed wikis from the wiki hosting service Wikia. Wikia is a hosting service for wikis with a focus fan sites for fiction franchises.[1] Its users are not charged for creating wikis, contributing or accessing information. Nevertheless, the operator Wikia Inc. is a for-profit company, and it generates profit from Wikia in the form of advertisement. In contrast to the broad scope of topics in the online encyclopedia Wikipedia, the main focus of Wikia is entertainment. Most Wikia communities cover

[1] In October 2016, Wikia.com has been renamed to “Fandom powered by Wikia” to strengthen the association with the “Fandom” brand.

| Wikia wiki | Revisions | Pages | Ratio |
|---|---|---|---|
| Disney | 158,733 | 1,710 | 92.82 |
| WoW | 122,449 | 1,148 | 106.66 |
| 24 | 56,509 | 914 | 61.826 |
| Tardis | 126,318 | 564 | 223.96 |
| Villains | 105,273 | 2,323 | 45.31 |
| The Walking Dead | 105,138 | 425 | 247.38 |
| Military | 75,028 | 13,189 | 5.68 |
| Wikipedia (sample) | 877,717 | 1,000 | 877.72 |

**Table 5.1:** Basic statistics of our data sets of seven selected Wikia communities. The Table states the full page and revision count, and the page-to-revision ratio.

topics from television, movie or (computer) game genres. Overall, Wikia hosts over 400,000 communities with over 200 million unique visitors per month.[1]

As opposed to Wikipedia, where the internal quality rating of articles follows a strict process, there are no global quality estimators for Wikia articles.[2] Since January 2012, Wikia provides a combined indicator of performance, traffic and growth for every individual community – the Wikia Activity Monitor (WAM).[3] This single score between 0 and 100 is recalculated on a daily basis, and is used to rank the communities. To prevent aimed manipulation of this score, the specific formula is not known to the public. As this score is applicable and comparable across Wikia communities, we used the WAM score as a global measure of community performance.

For our experiments, we chose a selection of seven English Wikia communities, based on high WAM score, reasonable size and genre diversity (see Table 5.1).

1 http://www.wikia.com/about

2 Although there are panels of contributors rating individual articles in some Wikia communities, there are no overarching norms for quality control across all Wikia wikis.

3 http://www.wikia.com/WAM/FAQ

More specifically, we excluded all Wikia communities that either

a) are non-english
b) have too unusual structure, like lyrics or answers
c) have over 200,000 revisions and would require very long computation time
d) did not have an available database dump from January 2016 or newer or
e) have a WAM score below 85

From the remaining choices, we selected the five communities with highest revision count: Disney (entertainment brand), Tardis (TV series), WoW (World of Warcraft – video game), Villains (evil characters from various media) and The Walking Dead (TV series). Since all five have very high WAM scores over 97, we handpicked two additional communities: "24" as an additional TV Series wiki with a WAM score of 86, and Military, which has a WAM score of 85, for additional genre diversity.

## 5.5 Our Approach

The following section provides an overview of our approach and explains essential principles. We applied automatic classification of revision categories (Section 5.5.1) and determined informal roles for contributors in writing communities (Section 5.5.2). We then created a graph based on individual contributor interaction and use a novel contributor role model to extract general collaboration patterns (Section 5.5.3). These patterns yielded insights about similarities and differences of wiki communities, and we explored the effect of these patterns on the success of a community. See Table 5.2 for an overview of our methods. To access and process data from Wikipedia and Wikia, we used the freely available Java Wikipedia Library [38]. Our results are based on the June 2015 database dump from the English Wikipedia and March 2016 dumps from

| **Method**, based on | **Involved Data Sets** | **Algorithm** | **Result** |
|---|---|---|---|
| **Revision** classification [33] | Training on labeled Wikipedia revisions, applied to unlabeled Wikipedia and Wikia revisions | Multi-label Classification | Revision categories |
| **Informal roles** [10] | Classified Wikipedia and Wikia revisions | K-means clustering | Wikipedia and Wikia informal roles |
| **Co-author network** [27] | Wikipedia sample, Individual Wikia communities | Sentence-level co-author network | Contributor interaction network |
| **Motif analysis (this work)** | Wikipedia interaction network, Wikia interaction networks | Motif analysis | Interaction chains; motifs |

**Table 5.2:** An overview of our Methodology.

Wikia. For the sake of understanding and readability, we present each method and result stepwise.

### 5.5.1 Revision Classification

We focused our research on contributor interaction in collaborative platforms based on writing processes. Contributors create online articles, and these articles are extended and refined by the same or other contributors. Every revision serves one or more purposes, such as adding content, spelling corrections or adding citations. Additionally, changes from one contributor can be completely revoked by another contributor. We classified revisions with the edit-based multi-label classification method proposed by Daxenberger and Gurevych [33] that has later been adapted to revision-level by Arazy et al. [10]. As training data, we used the data set described by the same article [10] with more than 13,000 manually labeled Wikipedia revisions and twelve revision types, such as “Add Substantive New Content”, “Rephrase Existing Text” or “Add Vandalism” (full list see Table 5.3). A detailed description of the revision types can be found in [10] and [6].

| | 24 | Disney | Military | Tardis | Villains | Walk. Dead | WoW | Wikia Average | Wiki-pedia |
|---|---|---|---|---|---|---|---|---|---|
| Add Citations | 1.18 | 1.11 | 2.22 | 1.47 | 0.54 | 1.12 | 2.90 | 1.51 | 1.84 |
| Add New Content | 21.99 | 20.59 | 7.03 | 19.94 | 10.73 | 23.36 | 22.33 | 18.00 | 22.29 |
| Add Wiki Markup | 26.93 | 30.49 | 36.31 | 27.72 | 36.46 | 24.91 | 27.78 | 30.09 | 28.09 |
| Create Articles | 0.52 | 0.25 | 6.32 | 0.18 | 0.72 | 0.06 | 0.32 | 1.20 | 0.42 |
| Delete Content | 11.16 | 8.60 | 3.25 | 9.49 | 4.16 | 10.63 | 10.64 | 8.28 | 6.64 |
| Fix Typo(s) | 12.88 | 11.46 | 22.10 | 16.28 | 10.88 | 12.96 | 11.93 | 14.07 | 12.23 |
| Reorganize | 9.82 | 13.21 | 12.43 | 10.14 | 28.42 | 7.44 | 9.25 | 12.96 | 4.59 |
| Rephrase | 4.43 | 5.40 | 1.16 | 5.22 | 2.96 | 6.09 | 4.38 | 4.23 | 3.88 |
| Add Vandalism | 8.80 | 6.45 | 1.18 | 6.79 | 4.10 | 9.03 | 7.72 | 6.30 | 9.93 |
| Delete Vandalism | 1.35 | 2.10 | 2.81 | 1.51 | 0.77 | 3.94 | 1.80 | 2.04 | 7.85 |
| Hyperlinks | 0.17 | 0.12 | 0.59 | 0.34 | 0.06 | 0.18 | 0.15 | 0.23 | 1.50 |
| Miscellaneous | 0.78 | 0.22 | 4.61 | 0.94 | 0.20 | 0.29 | 0.80 | 1.12 | 0.75 |

**Table 5.3:** Revision type distribution of different wiki communities, in percent.

We used this training data in a machine learning setup to create a model for automatic prediction of revision types on unseen data. Following Arazy et al. [10], we used a set of manually crafted features based on grammatical information (e.g. the number of spelling errors introduced or deleted), meta data (e.g. whether the author of the revision is registered), character- and word-level information (e.g. the number of words introduced or deleted) and wiki markup (e.g. the number of internal links introduced or deleted) of each revision. This information was then used by a Random k-Labelsets classifier [81], an ensemble method which optimizes the output of several decision tree classifiers, to classify revisions. The proposed method yields state-of-the-art performance on the Wikipedia dataset from Arazy et al. [10].

We applied the classification to a large number of Wikipedia and Wikia revisions, as listed in Table 5.1. The performance of this classification of Wikipedia revisions has been shown previously [10], but we transferred this method to Wikia revisions, using the same training data from Wikipedia. As we did not know about the effect of this

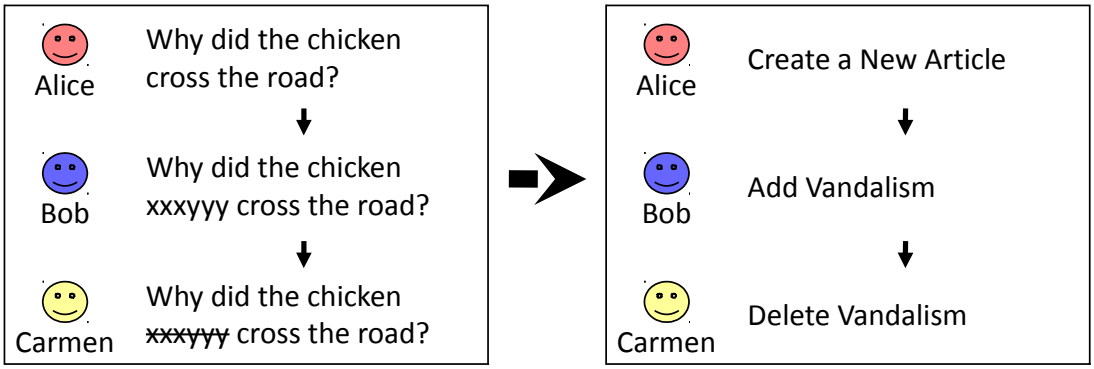



**Figure 5.1:** Example for Revision Classification. A model trained on Wikipedia data is used to classify Wikia revisions to one of twelve revision types.

change of domain (training on Wikipedia revisions, testing on Wikia revisions), we conducted a small-scale manual evaluation on Wikia data. Based on a manual evaluation of 100 random revisions from our Wikia sample, the classification of Wikia revision yielded results comparable to Wikipedia revision classification with 0.66 macro-F1 as compared to 0.68 in Wikipedia as reported by Arazy et al. [10]. Figure 5.1 illustrates the classification task on three example contributions.

### Results

Table 5.3 shows the distribution of the twelve revision classes on the Wikipedia sample and the seven Wikia communities. The distribution allows first observations on similarities and differences of wiki communities. Compared to the averaged Wikia communities, the Wikipedia data set has a 58% higher share of "Add Vandalism" revision, and 285% more "Delete Vandalism" revisions. Since Wikipedia attracts a much larger and broader audience as compared to Wikia, it also attracts more misbehavior, which results in the need of explicit counter-measures against these destructive actions. This

structural difference in contributor behavior is reflected in the increased "Add Vandalism" and, more strikingly, "Delete Vandalism" percentages.

Comparing the values of the seven Wikia communities shows their heterogeneous nature. The Wikia communities with a higher revision-to-page ratio, like Walking Dead, Tardis and WoW, are quite similar to each other and to the Wikipedia data. In contrast, Villains and Military, which both have a very low revision-to-page ratio, show significant differences.

The rate of "Reorganize Existing Text" revisions in Villains is more than twice as high as in every other data set. Military has an exceptionally high share of "Create a New Article" revisions, which is reasonable, given that it has by far the lowest amount of revisions per article. Furthermore, it seems to attract a high proportion of unusual edits, as shown by the above-average number of "Miscellaneous" revisions. Our findings indicate that maturity (as measured by the number of edits, as well as by age) influences revision behavior in online writing communities. Motivated by this finding, in the next section we go one step further and turn the revision behavior of individual contributors into a set of roles, which characterize the writing process in the entire community.

### 5.5.2 Informal Roles

In order to define generic motifs interaction (rather than individual editor collaborations), the individual contributors of both collaborative platforms – Wikipedia and Wikia – had to be mapped to a fixed set of roles that are based on revision types. Contributors with similar writing behavior in the context of a specific community should be assigned to the same informal role. We created revision type vectors for every contributor in every article, using the results of our revision type classification. Each vector contains the revision type frequency of every revision the contributor created for a given article,

normalized to sum up to 1. We detected informal roles from all vectors through a k-means clustering algorithm[1], with the number of clusters $k$ varying between 2 and 10. We compared the results via Overall Cluster Quality (OCQ) values, which is a balanced combination value of cluster compactness and cluster separation [10, 54]. The clustering with best OCQ values was chosen as the best informal role representation. For the 1,000 Wikipedia articles sample, this results in seven roles as described in previous work by Arazy et al. [10].

For our Wikia data sets, we considered two different approaches to cluster the contributors. The first strategy involves individual clustering of every Wikia community. As for Wikipedia, we used the same k-means clustering approach. From these possible clusterings, we selected the best option based on OCQ values. With this method, we were not able to create comparable informal roles across multiple Wikia communities (nor comparable to the ones we found for Wikipedia), which made it very hard to detect general collaboration patterns. The results can still be useful to compare Wikia communities, but the clusters from different Wikias are too diverse to draw meaningful conclusions across community borders, which makes this first approach unfeasible.[2] Therefore, we decided to map all Wikia contributors to a single, shared set of common roles, based on one global clustering on a combined data set of all Wikia revisions. We expect that a meaningful global role mapping for many Wikia communities might require a different number of clusters than Wikipedia. We considered all possibilities between 2 and 15 clusters. From these options, we selected the final clustering for all Wikia communities based on optimal OCQ values.

[1] Following Arazy et al. [10], we used the k-means++ method [14] as initialization for the clusters and tested a range of random seeds.

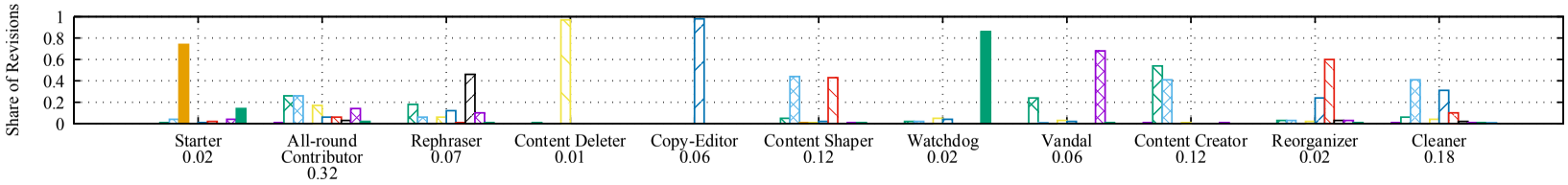
[2] Please note that this finding adds to previous work, which found that the nature of informal roles in Wikipedia remains stable over time [10]. Our results indicate that stable informal roles do exist within communities, this is not necessarily the case across different communities.

## Results

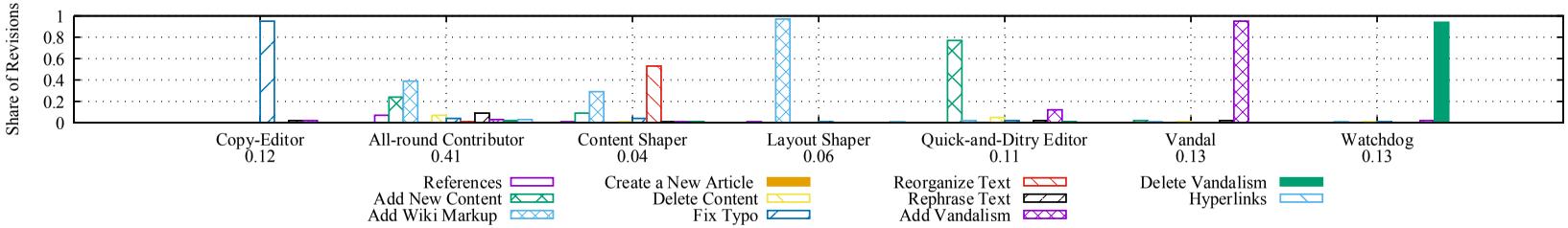
The best clustering for Wikia is presented in Figure 5.2a. It contains eleven informal roles: *Starter* (focus on creating new articles), *All-round Contributor* (no particular focus), *Rephraser* (focus on rephrasing content and adding text), *Content Deleter* (focus on deleting content), *Copy-Editor* (focus on fixing typos), *Content-Shaper* (focus on organizing content and markup), *Watchdog* (focus on vandalism detection), *Vandal* (focus on adding vandalism), *Content Creator* (focus on adding content and markup), *Reorganizer* (focus on moving text and fixing typos), and *Cleaner* (focus on fixing typos and markup).

A comparison to the Wikipedia cluster centroids discovered by Arazy et al. [10] (see Figure 5.2b) reveals some similarities, but also characteristic differences. One key similarity can be observed in the biggest cluster of each respective data set. These clusters both have a strong "All-round"-character, as their class distribution vectors have no clearly dominant dimension. This indicates that the majority of contributors does not focus on one single type of task.

The Wikia clustering contains several distinctive roles. Among these is the "*Starter*" role with a very large share of "New Article" revisions. Many Wikia articles only attract few edits after their creation, so an informal role that is limited to the creation of new articles is more likely. Communities with comparatively low number of revisions (see revision to page ratios in Table 5.1) – like "Military" – always contain an informal role with a strong focus on "New Article" revisions. The "*Content Deleter*" role is also unique to the context of Wikia, and contains contributors almost exclusively shortening and deleting content. Furthermore, we detected a role with contributors focusing on both




**(a)** Wikia informal roles.




**(b)** Wikipedia informal roles.

**Figure 5.2:** Global distributions of informal roles (fraction of contributors per cluster below role names) for our samples.

adding markup and fixing typos. Its scope is a bit broader as compared to the "*Spelling Corrector*" role in Wikipedia.

### 5.5.3 Collaboration Motifs

Having identified the different roles played by Wikipedians and Wikia contributors, we can analyze the interactions between types of contributors. Therefore, we used article-based co-author graphs, in which contributors form nodes and interactions between contributors form directed edges [27]. We calculated such a graph for each article from our Wikia sample. We mapped all contributors to the informal role they played in a particular article. We then counted all interactions, across all articles, between the different contributors. Lastly, we analyzed the effect of general interaction motifs on community performance (in Wikia). In the following, we describe this process in more detail.

#### Co-author Graphs

Brandes et al. [27] proposed an edit network based on sentence-level interaction. In their network, each Wikipedia article forms a graph $G = (V, E)$. The nodes $V$ correspond to the contributors who have performed at least one revision. The directed edges $E \subseteq V \times V$ represent interaction between a pair of contributors. As our intention is to understand collaboration between contributors based on their informal roles, we decided to slightly simplify the original co-author network of Brandes et al. [27]. In contrast to Brandes et al. [27], we defined the following types of interaction for a pair of contributors $u, v \in V \times V$: a) $u$ **supports** $v$, and b) $u$ **deletes** $v$. The support interaction indicates that contributor $u$ changed or added information to a sentence that contributor $v$ has created or edited. If contributor $u$ completely removes a sentence written by contribu-

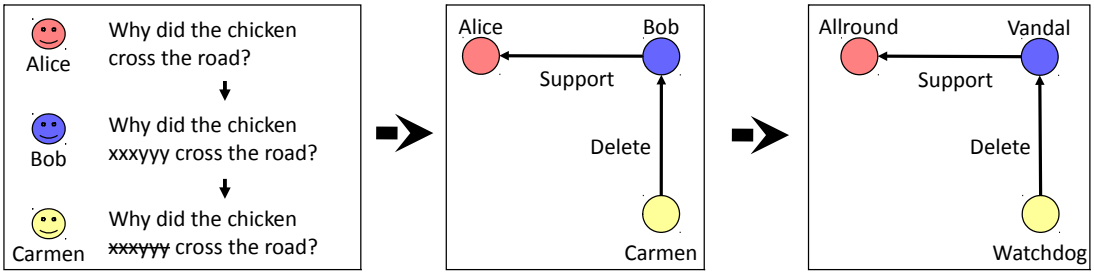



**Figure 5.3:** Example for co-author graph creation. First step: Identify sentence-level edits. Second step: Create graph with support and delete interactions. Third step: Replace contributor identification with respective informal role.

tor $v$ or reverts that contribution, we create a delete interaction. After the full graph is created, we replaced the labels of all nodes $V$ – the individual contributor – with their respective informal role, according to Section 5.5.2. See Figure 5.3 for a small example.

### Motifs

Based on the simplified co-author graph, we identified recurring collaboration patterns in the graph – our motifs [61]. In the context of this study, these motifs were defined as repeated interactions of the same type within the same edit context. As the delete interactions cannot be repeated within the same context – contributor $v$ adds or edits content, contributor $u$ reverts it – delete interactions are already interaction chains of maximum length. In contrast, support interactions can form chains of any length. If, for example, contributor $v$ adds content and contributor $u$ edits some of it and adds more information, the resulting interaction chain would be: $u$ supports $v$. Then, a third contributor $w$ adds some wiki markup in the same context, the resulting interaction chain would be: $w$ supports $u$ and $v$. To identify the basic motifs of the rather long interaction chains, they are split into pairs. As the example in Figure 5.4 indicates, the interaction chain "*All-round Contributor* supports *Starter* and *Copy-Editor*" is split

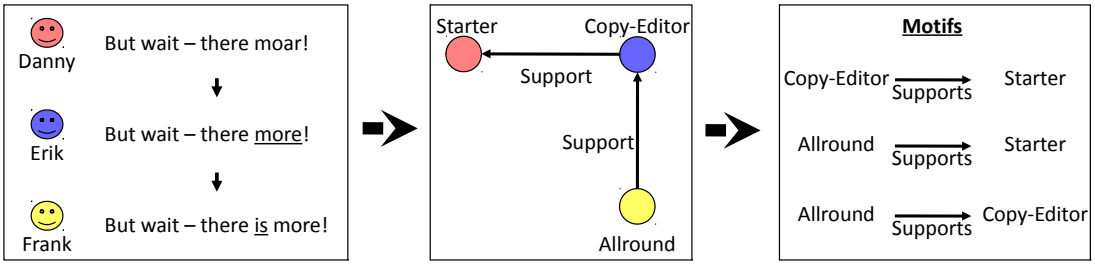



**Figure 5.4:** Interaction chains and pairwise motifs: a *Copy-Editor* supports a *Starter*; an *All-round contributor* further expands the work of both previous editors. This results in two additional support motifs.

into the two motifs "*All-round Contributor* supports *Starter*" and "*All-round Contributor* supports *Copy-Editor*". We consider these interactions motifs to be the building blocks of collaborative interaction.

We identified motifs of noticeable high frequency by comparison to randomly generated null-models [47], based on the interaction chains of informal roles. To generate a null-model, we kept the length and frequency of all interaction chains, but removed its informal role labels. This gave us the distribution of informal roles and a basic structure of interactions, with support chains of different length and frequency and a number of delete interaction pairs. We then redistributed the informal roles randomly to this structure. In this manner, we obtain the exact same chain lengths, same distribution of support and delete actions, and the same distribution of roles, but potentially different motifs. Figure 5.5 displays an example of the null-model creation process with two support chains, one delete chain and three different informal roles.

We created 1000 random null-models for every collaborative community. Based on these, we calculated the z-score of every support and delete motif as $z = \frac{F_G(G') - \mu_R(G')}{\sigma_R(G')}$, where $F_G(G')$ is the frequency of a given motif in our data. $\mu_R(G')$ and $\sigma_R(G')$ indicate mean frequency and standard deviation of that motif in the randomly generated null-models. The z-score compares one value of a group of values to the mean [48]. In our

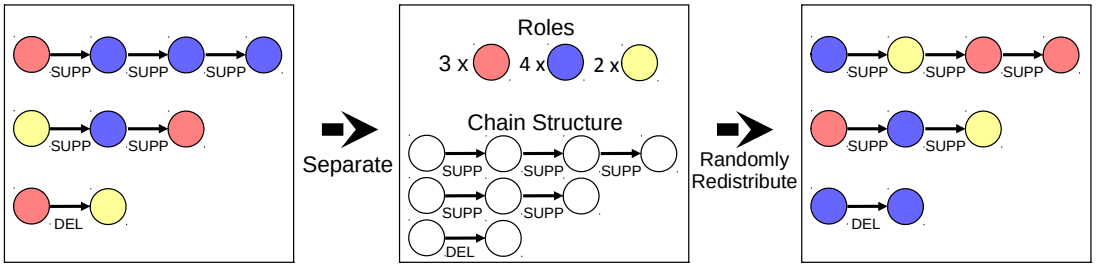



**Figure 5.5:** Null-model creation: we remove the role labels from all interaction chains and redistribute the labels randomly for each null-model.

case, high z-score values imply a remarkably high count of a particular motif compared to random chance. For example, the motif "*Rephraser* supports *Copy-Editor*" was found 5566 times in our Disney Wikia snapshot of January 2016. In our 1000 random null-models, the mean frequency of this motif was 4783.08 with a standard deviation of 43.69, which results in a z-score of 16.02. Since this is a relatively large positive number, it indicates that this motif is much more frequent in our real data in comparison to the random models.

Finally, we wanted to analyze the effect of unusual high or low frequency of specific motifs on the overall performance of the community. We conducted this analysis for every Wikia community, and used the Wikia WAM-score as an indicator of community performance. Since Wikia started publishing WAM-scores in January 2012, we considered seven points in time for our experiments in a 6-month rhythm, from January 2012 to January 2016. We determined the correlation between motif z-scores and the respective WAM score at each point in time with the Pearson correlation coefficient. In our case, a correlation coefficient of 1 for a specific motif would mean that a linear increase in the WAM score corresponds to a linear increase of the z-score of the motif. A correlation coefficient of $-1$ indicates linear negative correlation, where a linear increase in the WAM score corresponds to a linear decrease of the motif's z-score. If

there is no correlation between z-score and WAM score, the correlation coefficient is 0. The critical absolute value of the correlation coefficient for this amount of data is 0.754 for $p < 0.05$, so values lower than $-0.754$ or higher than 0.754 are considered to be statistically significant.

### Results

Our motif research is based on interaction graphs that contain all support or delete interactions between contributors in a single article. Figures 5.6 and 5.7 features visualizations of two prototypical graphs from one Wikia and one Wikipedia article. As seen in the graphs, the collaborative writing process in Wikia (see Figure 5.6) is much more centralized, as most of the interaction involves a small group of main contributors. These central persons can also be seen in the Wikipedia article graph (see Figure 5.7), but there are also small teams and subgroups that do not necessarily involve the main contributors. This difference is indicated by the Louvain modularity measure [21]. In the given example, the modularity of the Wikipedia graph is five times as high as the modularity of the Wikia graph (0.519 versus 0.101).

To identify and illustrate the most important motifs in our Wikia communities, we combined the significant positive and negative correlations across all Wikia communities into a single heat map. Figure 5.8 depicts the results for both support and delete interaction motifs. Light color (up to white) indicates positive correlation of the respective motif and the Wikia WAM score, dark color (up to black) indicates negative correlation. The support heat map shows a strong positive effect of the "*Rephraser* supports *Copy-Editor*" motif. Support interactions of similar roles are also positively correlated with the Wikia WAM score, like "*Reorganizer* supports *Content Shaper*" or "*Copy-Editor* supports *Reorganizer*". All these roles focus on small corrections and qual-

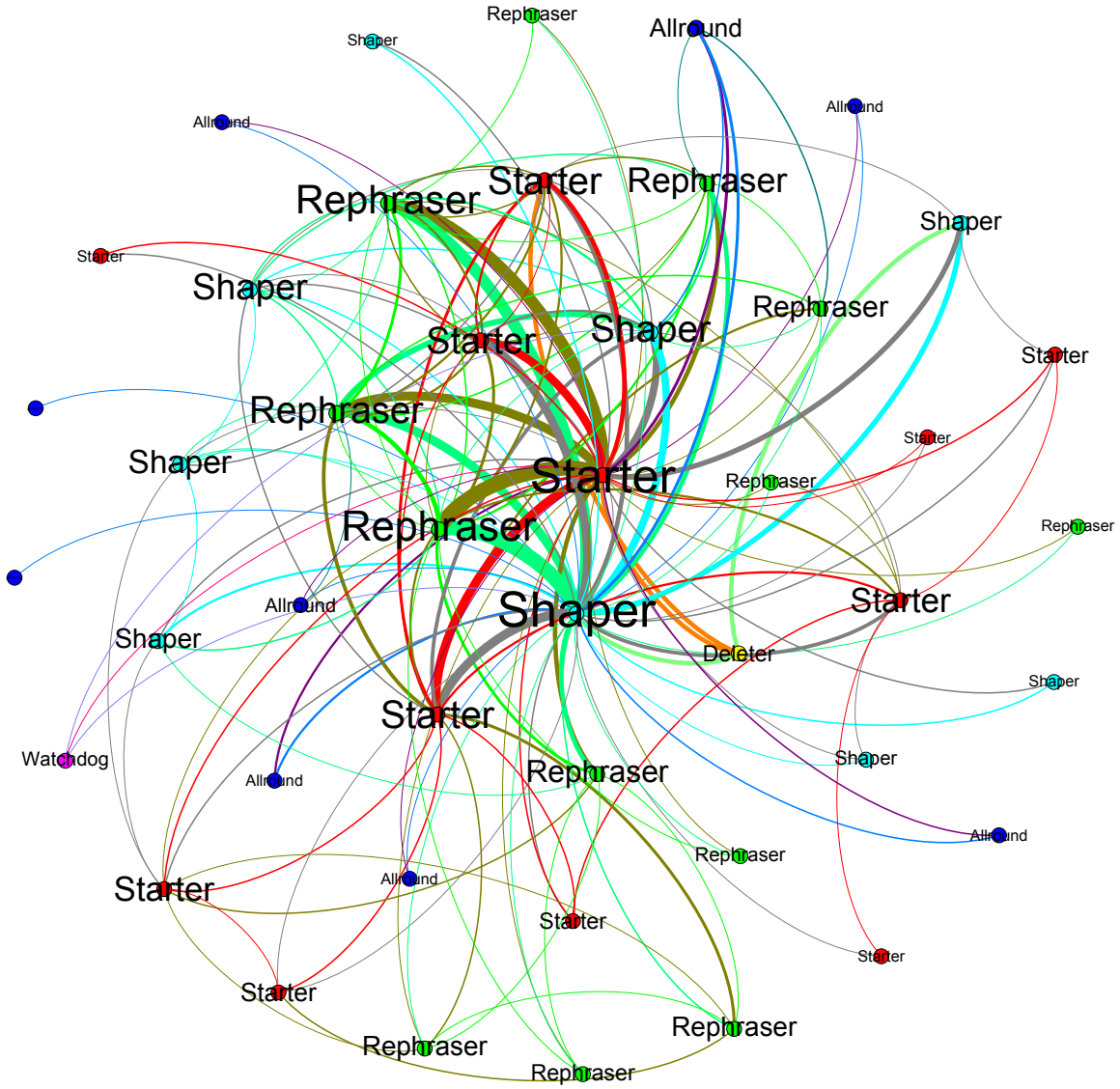



**Figure 5.6:** Graph visualization of Support interactions in the Disney Wikia article 'R2D2'. The nodes are single contributors, projected to their informal role. Edges indicate interaction between contributors. The node and edge sizes reflect the number of contributions and frequency of interaction.

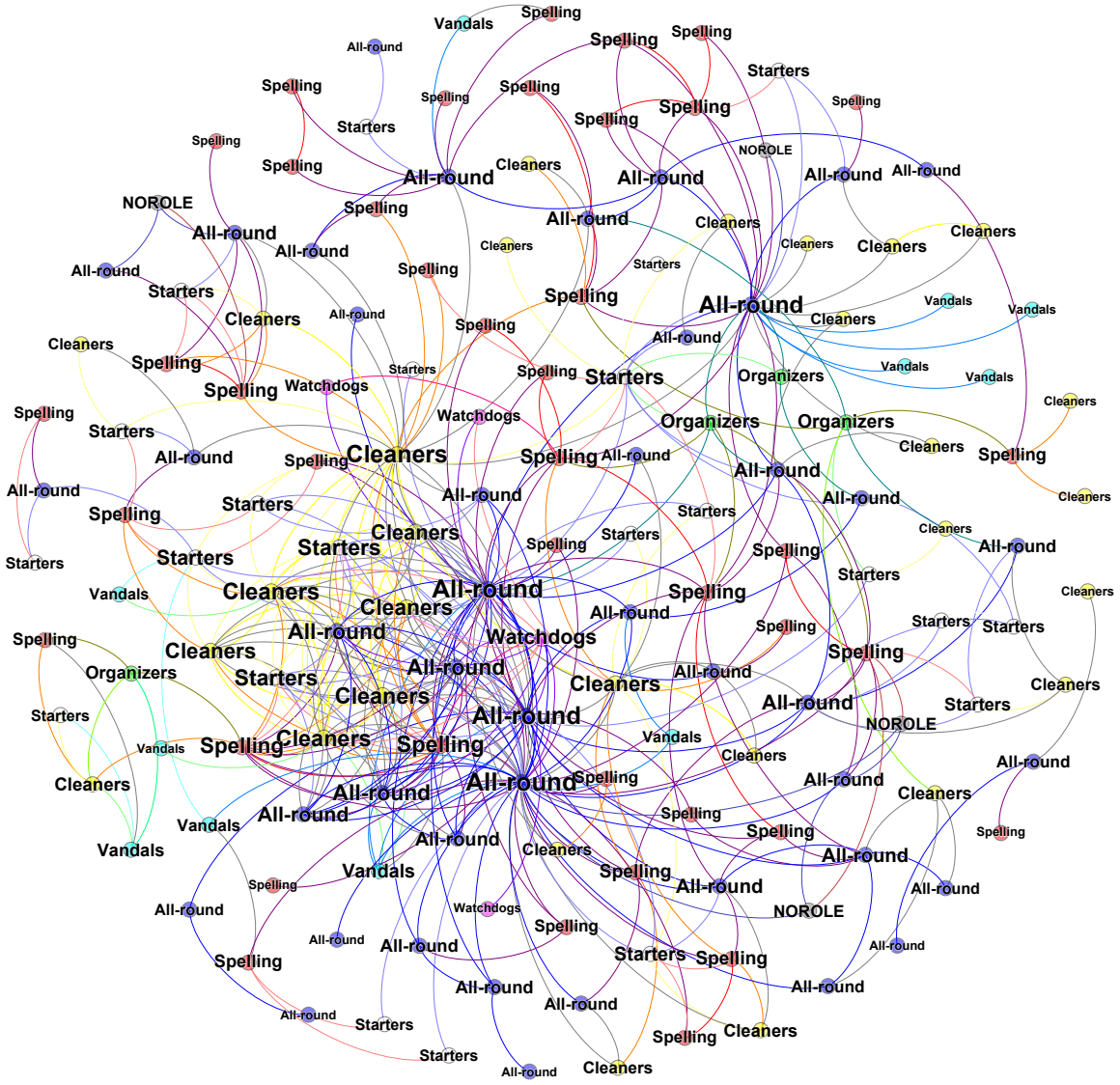
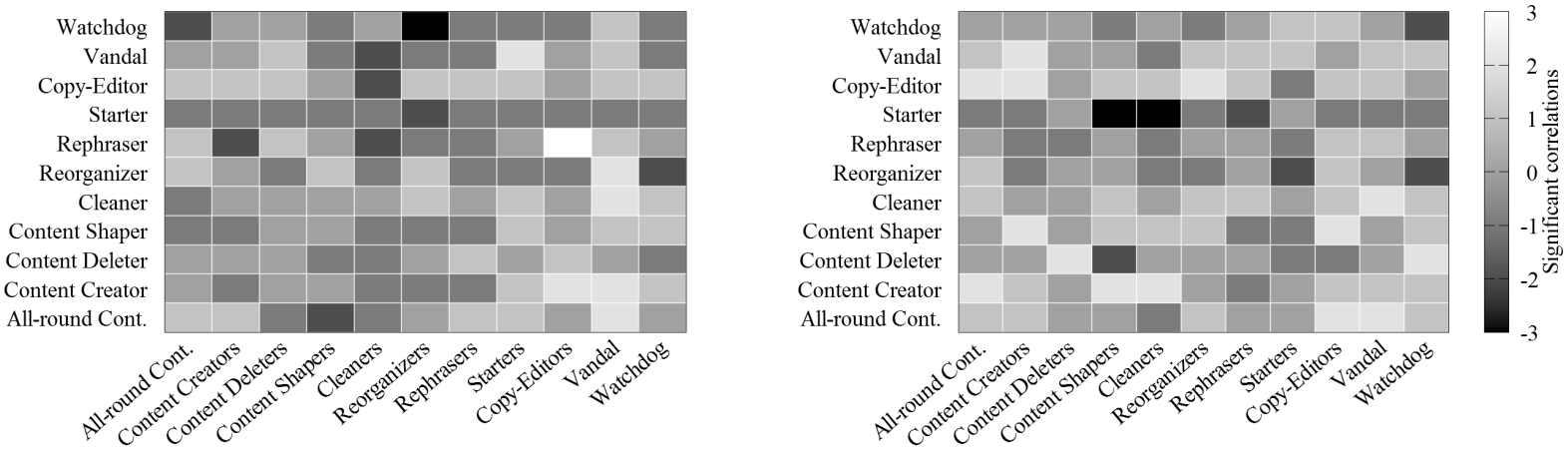



**Figure 5.7:** Graph visualization of Support interactions in the Wikipedia article 'Abscess'. The nodes are single contributors, projected to their informal role. Edges indicate interaction between contributors. The node and edge sizes reflect the number of contributions and frequency of interaction.




**(a)** Support motifs.

**(b)** Delete motifs.

**Figure 5.8:** Heatmap of correlation between motifs and Wikia WAM score. Light / dark color indicates a number of Wikia communities that showed statistically significant positive / negative correlations of the motif and Wikia WAM Score.

ity improvements, rather than the creation of new content. In contrast, the "*Content Creator* supports *Content Creator*" interactions shows slightly negative effects on the success of the community. The *Content Creators* role includes contributors that mostly focus on adding more content. This is an additional indication for the importance of quality improvements over quantity improvements.

The support motif "*Watchdog* supports *Reorganizer*" has the highest negative correlation with the Wikia WAM score. Almost all support interactions of the "*Watchdog*" role have negative values in the heat map. In contrast, the delete motifs heatmap show that delete interactions of "*Watchdog*" have more positive effects, which confirms that the main focus of this informal group should be on removing potentially problematic content. Delete interactions targeting the "*Vandal*" role are strongly correlated to high community success. All support and delete motifs from the "*Starter*" role have negative correlation coefficients.

## 5.6 Discussion

Arazy et al. [10] showed that the nature of informal roles, i.e. the result of clustering contributors, in Wikipedia did not differ much across two periods of time. They conclude that the set of informal roles they discovered shows a high stability within the online community Wikipedia. However, when comparing communities in Wikia, we found that the nature and maturity of a writing community might well have an influence on informal roles, and consequently, contributor interaction. The differences could be the result of the fact that Wikia is less restrictive with regard to its content. For example, Wikipedia follows the principle of the "Five Pillars"[1], whereas wikis on Wikia are not bound to an encyclopedic content and format. Wikipedia's principles offer a "boundary

[1] https://en.wikipedia.org/wiki/Wikipedia:Five_pillars

infrastructure" [10] which Wikia lacks. A lack of such collaborative principles results in more nuanced and less stable collaborative structures, indicated by the significant differences between individual informal role clusterings in Wikia and Wikipedia. As exemplified by the graph in Figure 5.6, Wikia articles tend to evolve around a central incubator, interacting with contributors working on the quality rather than the content of the article.

Our main findings connect motifs to community performance of Wikia platforms. The heatmaps in Figure 5.8 indicate that certain interactions between contributors have a significant impact on the overall performance of the community. To verify that the interaction of contributors is indeed key to success (or failure), we also tested the correlation of occurrences of revision types and informal roles with the Wikia WAM score, using the same dataset as in our motif experiments. As indicated in Table 5.4, the revision types "*Delete Substantive Content*" and "*Fix Typo(s) / Grammatical Errors*" have the greatest effect on the WAM score of all Wikia communities. As for informal roles, "*Rephraser*" and "*Cleaner*" show high positive and negative correlation coefficients. However, looking at the most significant interaction motifs, we see that they both show higher mean corre-

| | | **Corr.** | **SD** |
|---|---|---|---|
| Revision Types | Delete Substantive Content | -0.40 | 0.57 |
| | Fix Typo(s)/Gramm. Errors | 0.39 | 0.52 |
| Informal Roles | Rephraser | -0.49 | 0.46 |
| | Cleaner | 0.42 | 0.59 |
| Motifs | Cleaner supports Vandal | 0.68 | 0.17 |
| | All-round Contr. supports Content Creator | 0.59 | 0.28 |

**Table 5.4:** Mean correlation coefficient (Corr.) and standard deviation (SD) of the revision types, informal roles, and motifs with highest absolute correlation to Wikia WAM score.

lation coefficients and lower standard deviations across the different Wikia platforms. This shows that interaction is a more reliable predictor of community performance as compared to mere editing behavior or informal roles.

Despite the fact that vandalistic behavior and edit wars are less frequent on Wikia, vandalism and its countermeasures also seem to negatively impact collaboration on this platform, as shown by the negative effect of support interactions of "*Watchdogs*". Although these contributors might be necessary in their duty against vandalistic behavior, their contributions also seem to be controversial and thus have a negative impact on the community as a whole.

Looking at the motifs in more detail, we did find generally more positive influence of roles focusing on smaller quality improvements such as formatting and fixing typos. This is most noticeable in the roles "*Copy-Editor*" and "*Cleaner*" that are often positively correlated with the Wikia WAM scores. In other words, successful Wikia communities tend to place more value on content quality instead of quantity. In contrast, interaction of informal roles that are more concerned with adding or removing content, like "*Starter*" and "*Content Deleter*", did show negative or neutral effects on the community. Our finding adds to the work of Daxenberger and Gurevych [32], who found that high quality articles in Wikipedia attract more "surface" edits rather than revisions dealing with content extension or modification. Thus, our study confirms that their finding is valid for collaborative online communities other than Wikipedia. As a consequence, wiki organizers and administrators should emphasize the importance of both diversity and interaction among contributors, and incorporate this in their internal structures and processes. A potential application which would benefit from this analysis is e.g. online team formation [5], where contributors with different information roles need to be brought together in the right way.

## 5.7 Conclusion and Outlook

In this work, we demonstrated that motif analysis can assess quality not only in purely text based graphs, but also in graphs based on meta levels of text quality. To this end, we combined measures of implicit coordination with those from contributor interaction to assess community performance, and analyzed contributors' informal roles on two popular wiki platforms, namely Wikipedia and Wikia. While informal roles help to estimate what contributors do, interaction motifs from co-author networks reveal who they are working with. Rather than using collaboration patterns to detect trends [47], we leverage informal roles to analyze the effect of interaction on community performance. This approach helped to identify collaboration motifs with consistent positive or negative effect, which is not possible when looking at editing behavior or informal roles in isolation. Our results reveal a particularly positive influence of contributors with a focus on small contributions for text quality improvement. This finding, in combination with the more diverse collaboration patterns we found in different Wikia wikis, points to a clear need for measures to increase implicit coordination and quality assurance in public wikis by bringing together the right people [5].

We see several directions for future work. First, it might be very helpful to get insights about contributors' motivation. Recent work [11] revealed that changes in the implicit coordination of contributors can be linked to different motivational orientations. This dimension is not part from this thesis. Also, we had to rely on the untransparent WAM score as an indicator for community performance. Future work might look into different measures, e.g. by assessing the quality of all articles in a wiki.

# 6 Dynamic Metamotifs of Local Text Changes

## 6.1 Introduction

Several studies demonstrated the usefulness of motif analysis and other graph based methods in a variety of different scenarios and graph types. Many graphs, in particular those based on social science, are dynamic networks that can change over time. The connections of Facebook users is a prominent example of a dynamic network. Graphs based on text corpora are typically static as they do normally not include the version history of the documents. If different versions of a document are available, the graph structures of each version can be seen as one snapshot of a dynamic graph that changes for each revision of the source text. This possibility exists in Wikipedia and other online writing communities, as these platforms usually include the full version history of their content. In most graph based research on these data types, the dynamic nature is neglected. Some analysis do consider multiple snapshots of one document, but regard each separate graph representation of these document versions in isolation.

Instead, the changing graph representations can be interpreted as an evolutionary process - each graph of a later version emerges from an earlier version, with some modifications. If we transfer this view to motif analysis, motifs can evolve into other motifs as a consequence of local graph changes. We show that a closer inspection of these motif transitions can yield new insights about motifs and their characteristics. In order to

capture local graph modifications due to text revisions, we propose a new type of motif called *egocentric metamotifs*.

We commence this chapter with a sketch of our contribution in Section 6.2. We then present an overview of relevant related work in Section 6.3, and a short introduction of the corpus for this experiment in Section 6.4. We introduce and explain our extensions of motif analysis, egocentric metamotifs, in Section 6.5, followed by the experimental setup in Section 6.6. Finally, we will explore the results in Section 6.7 and close this chapter with a conclusion and discussion of our findings in Section 6.8.

## 6.2 Our Contribution

We have shown that analysis of local graph motifs can bring fruitful insights about quality in encyclopedic online communities - both in the form of direct text quality and overall community performance. We discovered that text contains distinctive patterns, which leads us to our next question: how do these patterns evolve over time? Are there certain types of motifs that change during the revision process of an article? Can we formalize characteristics of motif dynamics due to text changes?

In order to explore and answer these questions, we will extend our analysis, and not only cover local motifs, but the interplay of motifs. We will see that there are motifs of motifs, or *metamotifs*, that enable new ways to explore the data, and improve our understanding of text quality. In particular, we define a new characteristic for metamotifs in dynamic networks, called *metamotif stability*. This characteristic quantifies the prominence and fluctuation of metamotifs in the process of changes in the text, and the respective graph structure. We will take a closer look on the stability of metamotifs in articles of various quality stages, and discuss connections between stability and other motif features.

## 6.3 Related Work

Several studies exist that deal with the analysis of complex network dynamics and motifs, mostly in the field of cell biology and social network structures. Yaveroğlu et al. [92] exploit graph based characteristics to classify various real world networks, including Facebook, enzyme metabolism, and trade networks. In their work, they compare networks of different time periods to reveal connections between overall graph statistics and real world phenomena. For instance, they show how the position of a country in a trade network changes during a crisis, and measure these changes with centrality metrics. The work of Braha and Bar-Yam [26] follows a similar approach, but combines node centrality with motif distributions in university email networks. The composition of motifs in their data fluctuates heavily from day to day, and their results hint to correlations between these alterations and the prominence of central nodes. In contrast to our work, both of these approaches do not consider local motif dynamics or the individual changes, but compare global motif signatures of full graphs over time.

The concept of metamotifs, or motifs of motifs, have only mainly been used in the context of biology. Hau et al. [46] presents software for biological databases that includes searching capabilities for motifs in sequential data. The software enables the user to define arrangements of motifs, which the author call metamotifs. They do not, however, discuss characteristics or advantages. Piipari et al. [67] show that metamotifs are useful in classification tasks. Similar to the previous entry, they use sequential DNA data. To our knowledge, our evidence for the benefits of metamotifs in natural language data, or graph based approaches in general, are a novelty.

We also could not find research on motif stability in changing networks, or other characteristics similar to our definition. AbouAssi et al. [1] use the term motif stability to

quantify intrinsic, chemical properties of molecular structures, but the intention of this usage is quite different. There, the stability is a characteristic that arises from the combinations and connections of atoms in isolated chemical motifs, and not from motif or network dynamics.

## 6.4 Data

This part of research reuses our German Wikipedia corpus to study motifs in the revision process of articles. It contains all 2,338 featured articles and a purely random sample of 33,295 non-featured articles of a German Wikipedia snapshot from June 2015. In contrast to the previous study, we want to focus on motif *changes* in the article revision process, not the full motif spectrum of every article version. Therefore, we are able to include all revisions for every article used, and are not limited to a small subset due to processing capacities.
For more details about the contents of this corpus, refer to Section 4.4.

## 6.5 Egocentric Metamotifs

The term *metamotif* has not been used consistently throughout the scientific literature. Therefore, we want to clearly define our usage and our intentions. In this experiment, we focus on the interplay of motif combinations - we want to find patterns in the patterns, or motifs in the motifs.
Similar to a motif, we define a metamotif as a connected graph. In contrast to a motif, metamotifs are considered to be a specific combination of its included motifs. Also, metamotifs are not of fixed size, but the size of its sub-component motifs are - typically - fixed to three or four nodes. The exact shape, size and attributes of metamotifs, and therefore the algorithmic way to find them in a given graph, varies based on the use

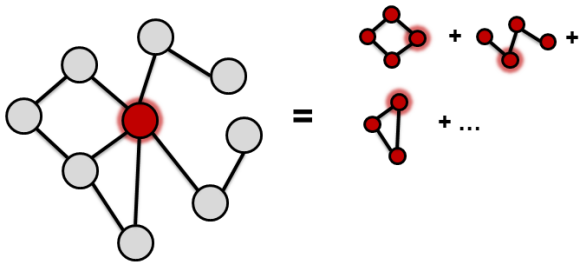


**Figure 6.1:** Explanation of egocentric metamotif. An egocentric metamotif encompasses all motifs in a small surrounding of a district, central node, or small group of connected nodes. In this example, the maximum distance of all connected nodes to the central node is two edges.

case and the context. Our metamotifs will be constructed around local text changes, so we will introduce the special type of egocentric metamotifs. An egocentric graph (also called egocentric network or ego-network) is a subgraph around a central node or small group of connected nodes that encompasses all nodes connected to the central node(s) up to a predefined maximum number of edges. We combine this notion of egocentric graphs with metamotifs: An egocentric metamotif is the specific combination of motifs in an egocentric graph, as illustrated in Figure 6.1.

## 6.6 Our Approach

We reused our graph definition from Section 4.5: The nodes of the graph are the sentences of the article version. Two nodes are connected by an edge if and only if these two conditions are fulfilled:

a) There exists at least one noun token that appears in both corresponding sentences.

b) The two sentences are separated by at most two other sentences in the document.

In the revision process of an article, most of the original text typically remains unchanged, as each edit process addresses a defined issue - e.g. add new content, fix grammar or update a specific part of the text. As a consequence, most of the motifs do

not change from one text version to the next. Changes in the motifs are only possible, albeit not mandatory, where there are changes in the underlying text. For that reason, we can use and exploit the local text changes for our metamotif search.

For every revision of a Wikipedia input article, we compared the text version before the revision and after the revision on the textual level. Whenever we detected differences, we built the egocentric graph according to our graph definition around the changed part in both versions of the text. The central node or group of nodes depends on the type of edit: If text was deleted, a set of nodes may exist in the first version, but not in the second version of the related graph. This set of nodes form the central nodes of the egocentric metamotif. In case of newly added text, new nodes might appear in the second version, and these form the central nodes. When text is changed, the central nodes are all nodes corresponding to changed tokens. The metamotif is then built around the central nodes in both article versions, including all connected nodes up to a maximum distance of two edges. We repeat this metamotif search for every textual difference in the revision, and for every difference, we determine the original metamotif and the target metamotif around the respective text edit. Figure 6.2 illustrates all three possibilities for local motif changes.

## 6.7 Results

To quantify the differences between metamotif changes, we introduce the novel concept of metamotif stability: The stability of a metamotif is the ratio of changes *INTO* the metamotif compared to changes *OUT OF* the metamotif (see Figure 6.3). In our interpretation, a metamotif with a high stability seems to be desirable, as it is be considered as a kind of “fixed point”. Metamotifs with low stability seem to be unwanted, as they are changed into another shape in most cases.

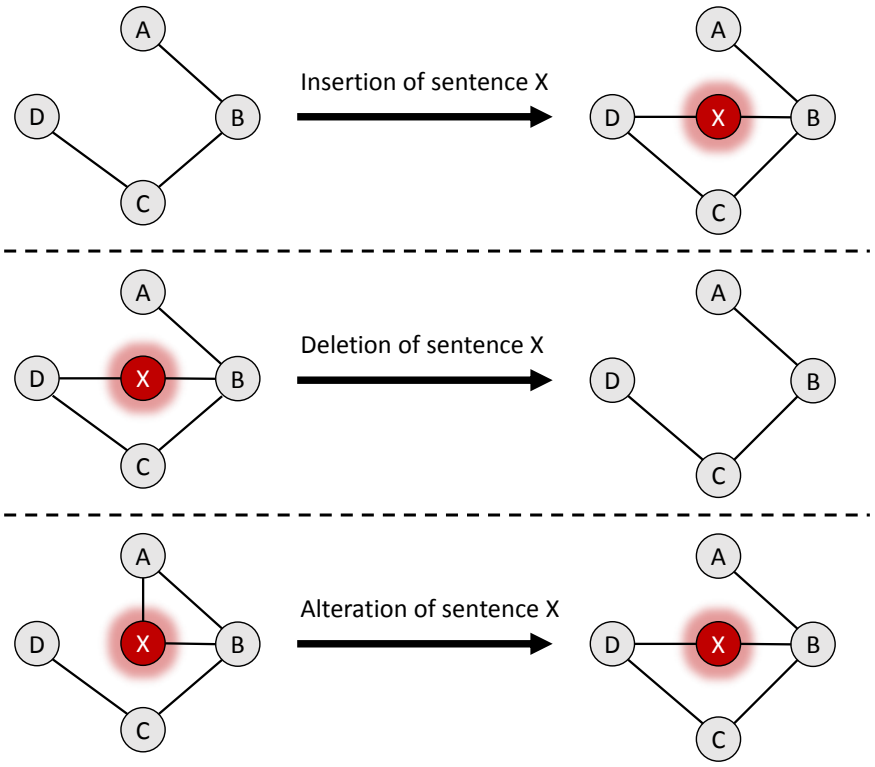



**Figure 6.2:** Possibilities for local motif changes. Motifs can change locally if sentences are added, deleted or edited. In each case, a metamotif is constructed around the changing node(s).


**Figure 6.3:** Explanation of metamotif stability. The stability is the ratio of changes INTO the metamotif compared to changes OUT OF the metamotif.

Looking at prominent metamotifs results, we suspected a correlation between stability and certain other parameters of the metamotifs, namely its size and its degree of connectivity. Metamotifs with a high stability seemed to be either relatively small, or fully (or almost fully) connected. This observation lead to the following hypothesis:

Let $S$ be that stability, $V$ the number of nodes, and $C$ the degree of connectivity of a given metamotif. $S$ is positively correlated with $\frac{C}{V}$.

We test this hypothesis with the Pearson correlation coefficient. To do that, we determine the stability of all metamotifs according to the formula:

$$Stability = \frac{Changes\_INTO\_metamotif}{Changes\_INTO\_metamotif + Changes\_OUT\_OF\_metamotif} \tag{6.1}$$

This value lies between 0 and 1. We consider metamotifs with $Stability > 0.5$ to have positive stability, and $Stability < 0.5$ to have negative stability.

The degree of connectivity can be measured in a number of ways that have similar results. We use the ratio of edges compared to maximum number of possible edges in the graph. In a directed graph $G = (V, E)$ with $|V|$ nodes and $|E|$ edges, the maximum number of edges is $|V| * (|V| - 1)$. In our graph representation, edges can only go from earlier sentences to later sentences, the direction of edges is fixed. As a consequence,

there are exactly half as many possible edges. Therefore, the ratio of edges to the maximum number of possible edges is:

$$Connectivity = \frac{|E|}{|V| \cdot (|V| - 1)/2} \tag{6.2}$$

Theoretically, the limits of this value in our graphs are 0 and 1. 0 indicates a disconnected graph with no edges, which is impossible in our definition of metamotifs, so the real minimum value is greater than 0. A maximally connected metamotif results in a connectivity of 1.

There are 5046 different metamotifs in our dataset, many of them appearing only once or twice. 69 metamotifs appear at least 100 times. Considering this subset of motifs, we get a Pearson correlation coefficient of 0.367, indicating a moderate positive correlation. If we include the 500 most prominent metamotifs, the correlation coefficient is still 0.206. Both results are statistically highly significant at $p < 0.001$.

We now look at the stability of metamotifs in different states of the articles' revision history. To do that, we split the Wikipedia revisions into three parts:

- NFA: All revisions of non-featured articles
- FA_EARLY: All revisions of featured articles before they got their featured status
- FA_LATE: All revisions of featured articles after they got their featured status

The stability of the most prominent metamotifs in the three subsets reveals interesting characteristics. Table 6.1 shows the mean stability and standard deviation of the 40 most prominent metamotifs per data set. The differences are small, but indicate a higher amount of stable metamotifs in higher quality articles. From the ten most prominent metamotifs of the NFA set, only two show positive stability. In the FA_EARLY set, five

| Data set | ∅ Stab. | SD |
|---|---|---|
| NFA | 0.496 | 0.071 |
| FA_EARLY | 0.506 | 0.061 |
| FA_LATE | 0.515 | 0.076 |

**Table 6.1:** Mean stability ($\varnothing Stab.$) and standard deviation (SD) of the 40 most prominent metamotifs per data set.

| Occurences | OUT | INTO | Stability | |
|---|---|---|---|---|
| 1074 | 455 | 619 | 0,576 | + |
| 603 | 271 | 332 | 0,550 | + |
| 423 | 196 | 227 | 0,536 | + |
| 375 | 171 | 204 | 0,544 | + |
| 313 | 146 | 167 | 0,533 | + |
| 307 | 148 | 159 | 0,517 | + |
| 265 | 143 | 122 | 0,460 | - |
| 225 | 134 | 91 | 0,404 | - |
| 183 | 107 | 76 | 0,415 | - |
| 162 | 80 | 82 | 0,506 | + |

**Table 6.2:** The ten most prominent metamotifs of the FA_LATE data set. The columns show the total number of metamotif occurrences, the number of changes OUT and INTO the metamotif, and the stability. The plus and minus signs indicate positive stability ($> 0.5$) or negative stability ($< 0.5$).

out of the top ten metamotifs have positive stability, and in the FA_LATE set, there are seven stable motifs - the six most prominent metamotifs are all stable. The details of this analysis are presented in Tables 6.2, 6.3 and 6.4.

## 6.8 Conclusion and Outlook

The metamotif analysis of dynamic Wikipedia articles has shown new insights about motifs and text quality. We introduced a novel definition of stability, and the results

| Occurences | OUT | INTO | Stability | |
|---|---|---|---|---|
| 1479 | 625 | 854 | 0,577 | + |
| 843 | 402 | 441 | 0,523 | + |
| 622 | 286 | 336 | 0,540 | + |
| 492 | 264 | 228 | 0,463 | - |
| 404 | 212 | 192 | 0,475 | - |
| 397 | 202 | 195 | 0,491 | - |
| 388 | 198 | 190 | 0,489 | - |
| 295 | 176 | 119 | 0,403 | - |
| 203 | 98 | 105 | 0,517 | + |
| 194 | 87 | 107 | 0,552 | + |

**Table 6.3:** The ten most prominent metamotifs of the FA_EARLY data set. The columns show the total number of metamotif occurrences, the number of changes OUT and INTO the metamotif, and the stability. The plus and minus signs indicate positive stability ($> 0.5$) or negative stability ($< 0.5$).

| Occurences | OUT | INTO | Stability | |
|---|---|---|---|---|
| 15905 | 6652 | 9253 | 0,581 | + |
| 12390 | 6261 | 6129 | 0,494 | - |
| 9311 | 4813 | 4498 | 0,483 | - |
| 7955 | 4088 | 3867 | 0,486 | - |
| 5829 | 3285 | 2544 | 0,436 | - |
| 4615 | 2254 | 2361 | 0,511 | + |
| 3944 | 2014 | 1930 | 0,489 | - |
| 3782 | 2259 | 1523 | 0,402 | - |
| 3734 | 2002 | 1732 | 0,463 | - |
| 2456 | 1471 | 985 | 0,401 | - |

**Table 6.4:** The ten most prominent metamotifs of the NFA data set. The columns show the total number of metamotif occurrences, the number of changes OUT and INTO the metamotif, and the stability. The plus and minus signs indicate positive stability ($> 0.5$) or negative stability ($< 0.5$).

indicate a significant correlation between stability, the size, and the degree of connectivity of a motif. In particular, smaller metamotifs tend to be more stable, as do fully connected metamotifs. In the context of our data set - encyclopedic Wikipedia articles - it makes sense that metamotifs should be small. The graphs of our texts are based on reused noun tokens, so smaller metamotifs indicate smaller units of meaning, distinct blocks of explanation, and less repetition. On the other hand, fully connected metamotifs also show higher stability, even if they have a fair number of nodes. We consider two possible explanations for this phenomenon: The extracted metamotifs are reoccurring boilerplate text, or consciously inserted repetitions. Looking into the actual metamotif data, we found the latter explanation to be true. Fully connected metamotifs of larger size mostly appeared in tables, listings and purposely inserted repetition.

In the revision process of higher quality articles, more stable metamotifs are added, and unstable metamotifs are reduced. This shows that metamotif changes are not random - they reveal additional insights about text quality. Using this knowledge to guide the editing actions of single users, or groups in a collaborative writing process, are interesting opportunities for future work.

# 7 Semantic Frame Metamotifs in Political Texts

## 7.1 Introduction

In social science, manual content analysis performed by human annotators is still the preferred and most widely-used method for empirical studies [50]. Due to high cost and effort, its applications are limited to small or medium scaled projects. Free access to open data and the influence of computational analysis across scientific disciplines has enabled many new branches of automated content research. In the realm of political analysis, valuable textual resources include online social platforms, but also collections and transcripts of political speeches, and a variety of political campaign material.

We apply motif analysis as an automatic analysis method on multilingual political data. This change of text type and genre serves as additional evidence for the possibilities of motifs. We assume that motifs can capture differences of politicians or political parties in terms of persuasion and lines of argumentation. For this reason, we use semantic frames for our graph representation instead of raw text. Semantic frames are related to concepts and topics, and serve as an abstraction layer. We also use motif and metamotif signatures in a machine learning classification task in comparison with other features.

In this chapter, we will first outline our contributions in Section 7.2, and discuss relevant literature on the topics of semantic frames and classification in political data in Section 7.3. We then present the newly created English and German corpora and their contents in Section 7.4, and explain the used semantic frame resources FrameNet [15]

and SALSA [35], together with the concept of semantic frames in Section 7.5. In Section 7.6, we illustrate our approach from plain text files to frame motifs, followed by both quantitative (see Section 7.7) and qualitative (see Section 7.8) evaluations of our results. Finally, we provide a summarizing conclusion in Section 7.9.

## 7.2 Our Contribution

In the final experiment chapter of this thesis, we discuss motif analysis and automated prediction techniques in the context of politics. We explore speeches of US politicians, German Bundestag debates and German party manifestos to find characteristic patterns in the respective language. Semantic frames from the FrameNet and SALSA knowledge bases serve as a generalization layer that allows us to create abstracted language graphs. We use motif signatures together with simpler features in a machine learning setup to predict party affiliation from source text of politicians. Here, our proposed metamotifs show a significant increase in discriminatory power compared to simpler methods. Finally, we take a closer look at prominent motifs, and discuss motif interpretability.
This analysis addresses the last main research question: Are metamotifs (motifs of motifs) an improvement over simple motifs and methods?

## 7.3 Related Work

There is extensive study on the content, structure, and uses cases of FrameNet. For a detailed overview, information about frame development and background on frame semantics, refer to [71]. Various applications of FrameNet and other lexical resources for several natural language processing tasks have been proposed, including work on machine translation [23], semantic role labeling [43] and natural language reasoning [64]. Other streams of literature focus on automatic extension and completion of these

knowledge bases. Shi and Mihalcea [77] combined the advantages of FrameNet with VerbNet and WordNet to a unified lexical resource, automatically extending the coverage of all individual data bases. Botschen et al. [25] used vector representations of FrameNet frames in trained neural networks for frame identification and prediction of missing frame-to-frame relations. In 2017, Peyrard et al. [66] evaluated frames in a combined scoring metric to judge summaries, together with a variety of other features and established measures.

In the last years, the availability of free textual resources has enabled a lot of social science research on automatic predictions of political topics. Bermingham and Smeaton [17] proposed a model of political sentiment in Twitter data, and combined sentiment analysis with supervised learning to successfully predict the result of an Irish election. In the same vein, Conover et al. [30] used Twitter data to predict the users' political alignment. Their results show that a model trained on hash tag metadata outperformed models that use full text. Due to the popularity of Twitter data in social science, Gayo-Avello [41] compiled a meta-analysis of political predictions in Twitter.

To the best of our knowledge, motif based approaches has not been applied on political data before. Motifs were applied on distantly related classification scenarios. For instance, Al Rozz and Menezes [4] discovered language motifs that could characterize authors from 100 books of the Gutenberg corpus. Their results show that graph motifs can be used as a signature for certain writing styles.

In our work, we hypothesize that motifs can not only signify personal writing or talking style, but may also indicate political affiliation.

## 7.4 Data

For our automatic analysis, we use three different text sources. We examine the language independence of our approach with the integration of both English and German material. Our English text corpus contains transcripts of official speeches from US presidents and presidential candidates, from George Washington in 1789 to the candidates of the most recent election in November 2016, Donald Trump and Hillary Clinton. The texts were extracted from the website of the American Presidency Project[1], a non-profit web archive of presidential documents. At the time of our experiments, January 2018, this yielded 33728 speeches from 105 different presidents and candidates. Table 7.1 shows an overview of the included documents over time.

For German data, we use texts from two different sources. First, we extract election manifestos of all German parties from the Manifesto Project[2]. According to the web-

| Years | Tokens | Documents |
|---|---|---|
| 1800 and earlier | 48183 | 18 |
| 1801 - 1850 | 431282 | 79 |
| 1851 - 1900 | 658592 | 102 |
| 1901 - 1950 | 811883 | 159 |
| 1951 - 2005 | 4860630 | 2061 |
| 2006 - 2009 | 6280831 | 7292 |
| 2010 - 2013 | 3419009 | 4446 |
| 2014 - 2017 | 4036405 | 5409 |

**Table 7.1:** Distribution of tokens and documents over time from the US presidential candidate corpus.

1 http://www.presidency.ucsb.edu
2 https://manifesto-project.wzb.eu/

site, it contains party policies of over 1000 international parties from 1945 until today. Since we only use the German segment of this collection, we obtain 95 party manifestos of 17 German parties, from all elections between 1949 and 2017. In contrast to these party election policies, we also use protocols from the German parliament as a textual resource of individual politicians. These plenary protocols were downloaded directly from the German Bundestag website [1], which offers plain text or XML files of all parliamentary sessions. Since we want to use this data set to get detailed qualitative insights about the specific language patterns, we concentrate on the protocols from the recently passed legislative period (October 2013 - October 2017), including all plenary sessions until February 2nd, 2018. This encompasses 258 protocols from 809 individual speakers. The protocols follow a strict format and every text segment is tagged with the name of the speaker. This makes the mapping of text segments to the individual speakers simple. Tables 7.2 and 7.3 show an overview of the two German corpora.

## 7.5 Semantic Frames

According to George Orwell, "Political language is designed to make lies sound truthful and murder respectable, and to give an appearance of solidity to pure wind."[2] In contrast to the encyclopedic language of Wikipedia, political language contains much more emotion and persuasion. Also, political debates of different time periods cover very diverse topics - from finance to health care, domestic and foreign affairs, and many more. To capture universal language and content patterns, we decided to use an abstraction

[1] https://www.bundestag.de/dokumente/protokolle/plenarprotokolle/plenarprotokolle

[2] Quotation by George Orwell, Available at: https://www.brainyquote.com/quotes/george_orwell_141761, Accessed: March 22, 2018.

| Party | Tokens | Documents |
|---|---|---|
| B90/Die Grünen | 354165 | 10 |
| FDP | 335306 | 19 |
| SPD | 284127 | 19 |
| CDU/CSU | 218356 | 19 |
| Die LINKE | 113547 | 3 |
| PDS | 47892 | 4 |
| Piraten | 34240 | 1 |
| AfD | 16893 | 2 |
| L-PDS | 8567 | 1 |
| DP | 3484 | 3 |
| KPD | 2327 | 1 |
| WAV | 2312 | 1 |
| DZ | 1551 | 2 |
| SSW | 1540 | 1 |
| GB/BHE | 677 | 1 |
| BP | 359 | 1 |
| DRP | 324 | 1 |

**Table 7.2:** Distribution of tokens and documents per party from the German Manifesto corpus.

| Party | Tokens | Speakers |
|---|---|---|
| CDU/CSU | 4069392 | 293 |
| SPD | 2726706 | 200 |
| B90/Die Grünen | 1957629 | 66 |
| Die LINKE | 1791749 | 74 |
| AfD | 48850 | 56 |
| FDP | 42221 | 40 |

**Table 7.3:** Distribution of tokens and individual speakers per party from the German Bundestag corpus. The corpus includes text of Bundestag debates from October 2017 until February 2018.

layer that generalizes over specific word usage. For that, we utilize the lexical semantic resources FrameNet [15] (for English language) and SALSA [35] (for German lanuage). FrameNet is a database that embodies the theory of Frame Semantics [39]: A frame is a prototypical situation with all involved participants. For instance, the FrameNet frame "Arrest" includes units of meaning that involve all aspects of an arresting action, like a suspect who gets arrested, the corresponding charge or the authority that conducts the arrest. FrameNet is a collection of over 1,200 semantic frames with over 200,000 manually annotated example sentences. Additionally, it contains pairwise frame-to-frame relations, like preceding frames or subframes.

The Saarbrücken Lexical Semantics Acquisition Project SALSA is a German project sponsored by the German Science Foundation DFG that uses FrameNet as a basis for its own semantic frame database. It contains large portions of the FrameNet frames, but also adds additional, manually annotated frame descriptions. These additional, SALSA specific "proto-frames" can be identified by their German name, whereas frames that are also covered in FrameNet inherit their English name.

## 7.6 Our Approach

In order to analyze motifs of frames in our text corpora, we use a frame identification system to map the extracted plain texts to semantic frames (see Section 7.6.1). We then create graph structures to extract frame motifs (see Section 7.6.2).

### 7.6.1 Frame Prediction

Mapping specific tokens to a matching frame is a non-trivial task. We use the system of Botschen et al. [25] that uses frame embeddings for prediction. In order to predict

the semantic frame of a single token, this system needs the token in sentence context, the lemmatized token with part-of-speech tags and dependency parsing tags, which we all provide with the Stanford CoreNLP natural language processing toolkit [57]. In this way, we get a frame prediction for every token in our source data with the exception of common function words and punctuation. Table 7.4 demonstrates the preprocessing and prediction annotations on an example sentence.

Semantic frames, as defined by the FrameNET and SALSA databases, are not bound to a specific text genre. Nevertheless, the prediction quality of this method on our political texts was unclear, as there is no comparable study or data set with manually annotated frames available. However, the quality of this prediction is crucial for every following computation step - frame motifs are meaningless if we cannot trust the frame annotations itself. To test the plausibility of the predictions, we conducted a small-scale manual evaluation on 100 random English and German sentences from our corpora, respectively. The frame prediction yielded an accuracy of 85.51% on the English sentences, and 78.76% on the German sentences. These values are comparable to the results by Botschen et al., who reported 88.66% average accuracy on English, 80.76% on German data.

### 7.6.2 Motif Extraction

We want to find local patterns in the political language that might be tied to persuasion and argumentation. Therefore, we decided to constrain the motif search within sentence limits. To do this, we transformed every frame prediction into a graph node, and connected two nodes if and only if the corresponding tokens are withing the same sentence, and no other predicted token lies in between. This yields node chains that emerge from single sentences. Figure 7.1 shows an example of this construction. Note

| Nr. | Token | Lemma | POS | DP head | DP label | Frame prediction |
|---|---|---|---|---|---|---|
| 1 | I | I | PRP | 2 | nsubj | |
| 2 | have | have | VBP | | | |
| 3 | a | a | DT | 4 | det | |
| 4 | dream | dream | NN | 2 | dobj | Cogitation |
| 5 | that | that | IN | 22 | mark | |
| 6 | one | one | CD | 7 | num | Cardinal_numbers |
| 7 | day | day | NN | 22 | tmod | Calendric_unit |
| 8 | the | the | DT | 9 | det | |
| 9 | sons | son | NNS | 22 | nsubj | Kinship |
| 10 | of | of | IN | 9 | prep | |
| 11 | former | former | JJ | 12 | amod | Time_vector |
| 12 | slaves | slave | NNS | 10 | pobj | Experiencer_focus |
| 13 | and | and | CC | 9 | cc | |
| 14 | the | the | DT | 15 | det | |
| 15 | sons | son | NNS | 9 | conj | Kinship |
| 16 | of | of | IN | 15 | prep | |
| 17 | former | former | JJ | 19 | amod | Time_vector |
| 18 | slave | slave | NN | 19 | nn | Kinship |
| 19 | owners | owner | NNS | 16 | pobj | Possession |
| 20 | will | will | MD | 22 | aux | Required_event |
| 21 | sit | sit | VB | 22 | xcomp | Posture |
| 22 | down | down | RP | 24 | prt | |
| 23 | together | together | RB | 24 | advmod | Collaboration |
| 24 | at | at | IN | 24 | prep | |
| 25 | a | a | DT | 29 | det | |
| 26 | table | table | NN | 27 | pobj | Containing |
| 27 | of | of | IN | 29 | prep | |
| 28 | brotherhood | brotherhood | NN | 30 | pobj | Organization |
| 29 | . | . | . | 2 | punct | |

**Table 7.4:** Example of preprocessing and frame predictions of an example sentence. Preprocessing includes the token with its lemmatized form, the part-of-speech (POS) tag, the ID of the dependency parsing (DP) head tokens and the dependency relation label. A semantic frame is predicted for every non-function word.

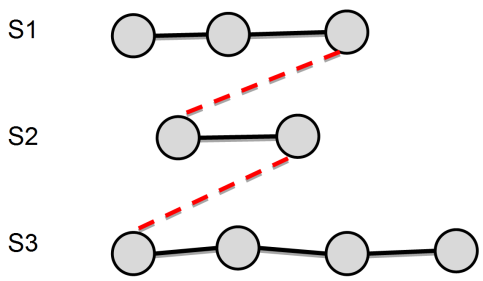



**Figure 7.1:** Example of semantic frame graph construction. Each chain of nodes consists of the predicted semantic frames of one sentence. The red edges between the end of one sentence and the beginning of the next sentence are used for metamotif search, since metamotifs can combine motifs of adjacent sentences.

that the last node of a sentence is connected with the first node of the following sentence via a special type of edge that is only used for searching meta-motifs, which will be explained later.

Using this model, every path of two or more nodes is considered to be a motif. We limited motif size to a maximum of four nodes to ensure motif interpretability. Similar to Chapter 4, we created motif signatures for every party / politician in our data sets by counting all extracted motifs from all texts of the respective origin and transforming these counts into a vector, normalized to a sum of one. Note that a vector from the English data set can be compared to any other English vector, in terms of distance or similarity. The same holds for vectors from the German data set. Since the underlying frame databases for English and German data differ, they can not be compared mathematically across languages. As a baseline comparison, we also created frame signatures in the same way, counting just the frame predictions, normalized to a sum of one.
Finally, we extracted metamotifs. Two motifs in our graph structure were considered to form a metamotif if and only if they are connected with an edge, or with a special type of edge that connects two consecutive sentences. In theory, this notion of metamotifs can be extended to encompass any number of connected motifs. Since interpretabil-

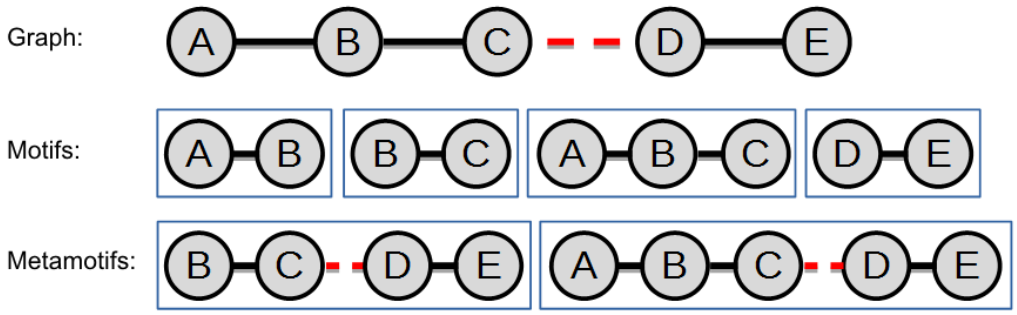



**Figure 7.2:** Example of motif and metamotif extraction. The graph is constructed from two sentences. The three frames "A", "B" and "C" are predicted from the contents of the first sentence, two frames "D" and "E" from the second sentence. In this graph, we find four motif occurrences - three motifs of two nodes, and one motif of three nodes. Also, we extract two metamotifs of two connected motifs. The metamotif connection is able - but not required - to cross sentences.

ity and computation effort are already problematic at two connected motifs, we did not consider metamotifs of bigger size. We again transformed the metamotif results to metamotif signatures. Figure 7.2 demonstrates the motif search on a small example.

## 7.7 Quantitative Results

Without further modifications, both motif and metamotif signatures were very high dimensional vectors of extreme sparsity. In the German data set, over 2,500,000 unique motifs and an exponentially higher number of unique metamotifs were found at least once. Most of the motifs and metamotifs are extremely rare, though, only appearing a few times in the whole data set. Figure 7.3 shows the frequency distribution of motifs. 2,254,217 motifs appear only once in the whole data set, 167,891 motifs appear twice, 45,314 motifs three times. Since we are not interested in frame combinations that are so rare, we decided to keep only the motifs that were used at least three times by any politician or party, and metamotifs that were used at least three times by any politician

or party. This filters out rare cases, and reduces the vector dimensions tremendously. In case of the German data, only 3283 unique motifs and 2902 unique metamotifs remain.

Do metamotifs have increased discriminatory power compared to motifs and simpler Methods? To address this main research question and quantify the discriminatory power of motifs and metamotifs, we use the frame, motif and metamotif signatures in a prediction scenario. As an additional baseline feature, we also compared against POS tag signatures. The experiments were conducted on the German Bundestag and the US presidency corpora. Each training example in this setup consists of the frame, motif or metamotif signature of a single politician, the goal is the prediction of the respective political party.

The political system of the United States is mainly a two-party system of the Democratic Party and the Republican Party, with only minor influences of third parties. Therefore, we concentrated on the distinction between Democratic and Republican candidates, and excluded candidates that ran for both the Democratic and the Republican party, or ran for presidency before those parties were established in the current form.

In the German Bundestag of October 2013 until October 2017, the following four parties were present: The coalition of Christian Unions CDU and CSU, the Social Democratic Party SPD, the Greens/Alliance90, and the left-wing party “Die LINKE”. Beginning with October 2017, two additional parties joined the newly formed Bundestag: The Free Democratic Party FDP and the Alternative for Germany AfD. In other words, in the Bundestag corpus up to October 2017, there are four possible output classes. In the whole corpus, or the subcorpus starting from October 2017, there are six possible output classes.

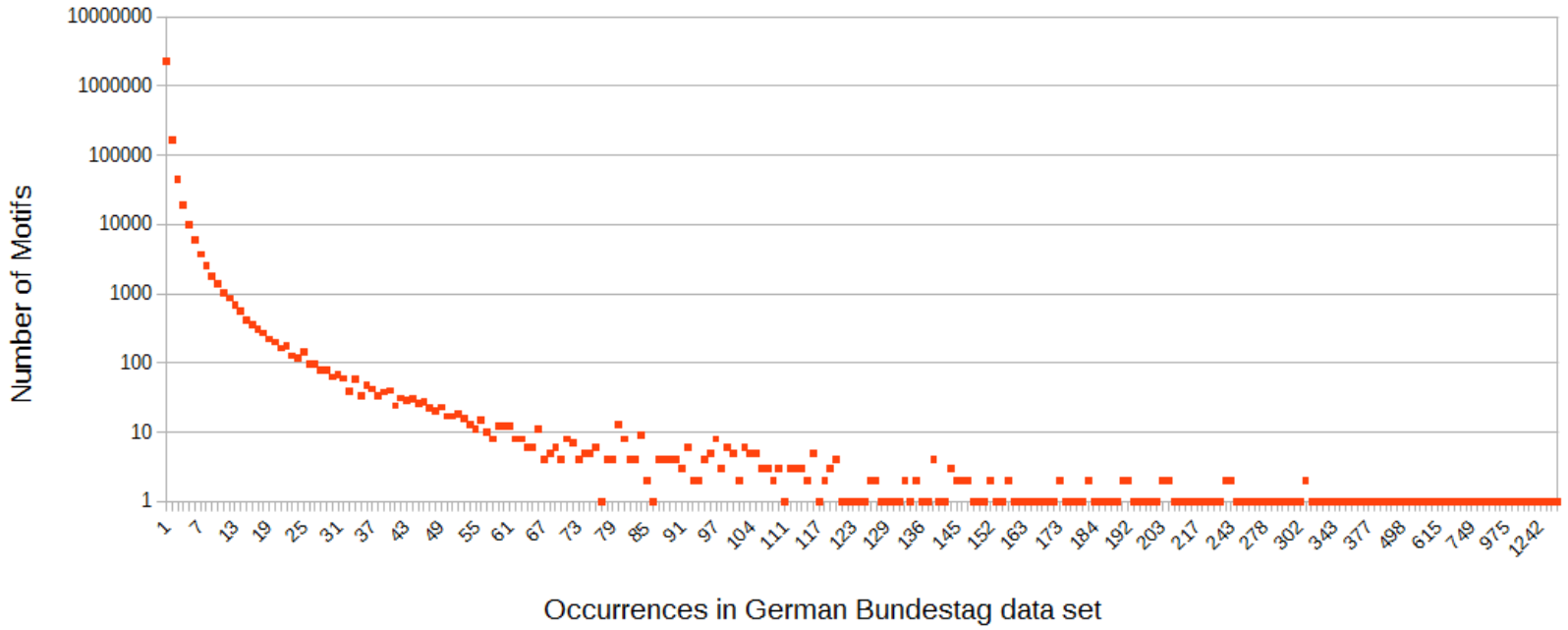



**Figure 7.3:** Motif frequency distribution in Bundestag data set. The x-axis denotes occurences of motifs, the respective y value is the number of motifs with this frequency. The y axis is of logarithmic scale.

We applied six different classifiers to the US data set and the three German data set segments - October 2013 to October 2017, October 2017 to February 2018, and the full corpus. Using the Weka machine learning toolkit [87], we ran three tree learning algorithms (J48, RandomTree and REPTree), two rule learners (OneR and PART) and the Bayes Network classifier BayesNet. Each experiment is repeated ten times with ten-fold cross validation. We also compare all classifiers against the baseline classifier ZeroR that always predicts the majority class. The averaged accuracy results and standard deviations are presented in Tables 7.5, 7.6 and 7.7.

| | ZeroR | J48 | RandomTree | REPTree | OneR | PART | BayesNet |
|---|---|---|---|---|---|---|---|
| POS | 36.262 | 31.792 ± 1.74 | 30.655 ± 1.05 | 36.590 ± 1.92 | 34.079 ± 0.71 | 32.793 ± 1.43 | 36.060 ± 0.76 |
| Frames | 36.262 | 30.940 ± 1.28 | 29.084 ± 1.51 | 34.158 ± 1.47 | 36.138 ± 1.25 | 30.198 ± 1.69 | 32.550 ± 0.48 |
| Motifs | 36.262 | 33.292 ± 1.65 | 26.114 ± 2.27 | 34.282 ± 0.78 | 33.910 ± 1.61 | 28.837 ± 0.97 | 36.015 ± 0.57 |
| Metamotifs | 36.262 | **35.068** ±1.32 | **33.705** ±1.81 | **37.376** ±1.43 | **37.252** ±0.68 | **35.440** ±1.70 | **39.623** ±0.89 |

**Table 7.5:** Accuracy results for party prediction of 809 German politicians in the full Bundestag data set (October 2013 - February 2018), with standard deviation over 10 cross validation evaluations.

| | ZeroR | J48 | RandomTree | REPTree | OneR | PART | BayesNet |
|---|---|---|---|---|---|---|---|
| POS | 42.622 | 33.568 ±1.32 | 37.031 ±1.13 | 41.942 ±0.94 | 37.115 ±1.36 | 35.087 ±1.72 | 45.290 ±0.70 |
| Frames | 42.622 | 34.724 ±1.85 | 36.661 ±1.46 | 41.132 ±1.21 | 38.599 ±1.88 | 31.445 ±1.94 | 40.685 ±1.36 |
| Motifs | 42.622 | 37.257 ±0.91 | 32.488 ±1.57 | 40.834 ±1.39 | 40.685 ±1.55 | 34.426 ±1.36 | 42.473 ±1.22 |
| Metamotifs | 42.622 | **38.599** ±0.62 | **38.450** ±0.75 | **42.473** ±0.34 | **43.815** ±0.26 | **36.065** ±0.94 | **47.093** ±0.67 |

**Table 7.6:** Accuracy results for party prediction of 671 German politicians in the Bundestag data set from October 2013 to October 2017, with standard deviation over 10 cross validation evaluations.

| | ZeroR | J48 | RandomTree | REPTree | OneR | PART | BayesNet |
|---|---|---|---|---|---|---|---|
| POS | 23.361 | 17.532 ±1.64 | 18.094 ±1.34 | 20.216 ±1.87 | 19.544 ±1.18 | 20.682 ±1.91 | **23.921** ±1.77 |
| Frames | 23.361 | 18.803 ±1.45 | 18.518 ±1.58 | 19.943 ±1.32 | 20.227 ±1.33 | **22.507** ±1.64 | 23.646 ±1.01 |
| Motifs | 23.361 | 17.094 ±1.57 | 15.669 ±1.82 | 23.646 ±1.10 | **23.361** ±1.29 | 15.384 ±1.52 | 23.722 ±1.46 |
| Metamotifs | 23.361 | **20.512** ±0.56 | **18.803** ±0.61 | **24.786** ±0.66 | 23.076 ±1.08 | 19.658 ±1.29 | 23.819 ±0.91 |

**Table 7.7:** Accuracy results for party prediction of 351 German politicians in the newly formed Bundestag data set from October 2017 to February 2018, with standard deviation over 10 cross validation evaluations.

On the full data set (see Table 7.5), regular motifs did not surpass simple frame counts, as they achieve an equal amount of lower and higher scores in comparison. However, metamotif signatures outperform POS tags and both motif and frame signatures in every experiment. All features did not compare well to the ZeroR baseline though. Only metamotifs in combination with the BayesNet classifier managed to yield statistically significant improvements to this baseline. This also holds true for the data set up to October 2017 (see Table 7.6) - metamotifs constantly achieve higher prediction accuracy than the other features, and the improvements of BayesNet are statistically significant. On the smallest subset, beginning with October 2017 (see Table 7.7), the results are not so clear. Metamotifs perform best in four out of six methods, outperformed once by both motifs and frames. All accuracy scores on this subset are much lower than on the other two data sets, though. This drop in performance is a consequence of much smaller amount of text and increased prediction difficulty, as there are now six possible party output values instead of four.

We tried to confirm the performance of metamotif features on the US data set. As shown in Table 7.8, metamotifs signatures performed best in every run, except for the Naive Bayes classifier. Generally, the results vary much more than in the German data set experiments from classifier to classifier. Also, the deviation between runs increased massively compared to previous experiments. We create one instance for every candi-

| | ZeroR | J48 | RandomTree | REPTree | OneR | PART | BayesNet |
|---|---|---|---|---|---|---|---|
| POS | 60.919 | 60.154 ± 3.12 | 59.943 ± 5.65 | 59.421 ± 3.81 | 53.313 ± 3.47 | 59.476 ± 4.04 | 55.641 ± 3.51 |
| Frames | 60.919 | 57.356 ± 4.28 | 59.885 ± 6.28 | 60.114 ± 2.76 | 52.528 ± 3.79 | 58.505 ± 4.31 | **59.425** ±3.38 |
| Motifs | 60.919 | 55.747 ± 6.06 | 57.701 ± 5.52 | 59.885 ± 3.26 | 52.068 ± 3.21 | 55.172 ± 4.97 | 54.827 ± 3.46 |
| Metamotifs | 60.919 | **62.758** ±2.43 | **61.264** ±2.76 | **61.609** ±4.66 | **57.931** ±3.80 | **61.954** ±3.08 | 53.563 ± 2.31 |

**Table 7.8:** Accuracy results for party prediction of 87 US presidency candidates, with standard deviation over 10 cross validation evaluations. The classification is a binary decision between Democratic and Republican party.

date of the Democratic or Republican party, so only 87 instances in total. With this few data, we expected the training process to be highly unstable. We can see a tendency that metamotif signatures performed better than every other comparing feature, but none of the improvements in this data set is statistically significant.

## 7.8 Qualitative Results

The large number of unique motifs and metamotifs eliminates the option for systematic inspection of the whole spectrum. For a qualitative evaluation, we therefore take a focused look at the most prominent motifs of individual politicians and parties, using all three data sets. Our observations of these two source types will be discussed separately. We hope to find common patterns in the overall political language that can generalize over topics and phrases. Also, we will discuss the feasibility of qualitative analysis on the more complex metamotifs.

### Motifs of Politicians

In general, the most prominent motifs in our two politician data sets across individuals are comprised of mostly forms of addressing and common phrases. Here are some instances for frequent motifs and exemplary occurrences in the data. The motif name is derived from its combination of semantic frames.

- Motif: Statement - Leadership

  Example: I don't think so, Mr. President!
- Motif: Operating a system - Leadership

  Example: That's why I'm running for President!

- Motif: Statement - Political Locale

  Example: What about this country?
- Motif: Desirability - Telling

  Example: Ich will noch eines sagen! (I want to say one more thing!)
- Motif: Experiencer subj - Collaboration - Collaboration

  Example: Liebe Kolleginnen und Kollegen. (Ladies and getlemen.)

These common phrases are not of topical nature. This is somewhat expected, because the most prominent motifs emerge from shared vocabulary, without individual themes or characteristics. To identify custom motifs of individual politicians, we determine the most frequent motifs of their respective documents, disregarding all motifs that are common over the whole data set. Here are some examples of individual motifs:

- Motif: Activity finish - Activity start

  Example: We'll finish what we started! (Barack Obama)
- Motif: Finish competition - Change of Leadership

  Example: We will win this election! (Hillary Clinton)
- Motif: Building - Architectural part

  Example: Yes, we will build a wall. (Donald Trump)
- Motif: bleiben - Statement

  Example: Nichts muss bleiben, wie es ist. (Anton Hofreiter)
- Motif: Request - Questioning

  Example: Ich darf bitten, auf die Frage zu achten. (Volker Beck)

As we see, some motifs of individual politicians did capture characteristic phrases, especially in the English data set, which contains public speeches. A large portion of the speeches are designed for election campaigns, where repetition of distinctive phrases is

a deliberate campaign strategy [36]. The German parliamentary debates do not have this goal, and individual prominent motifs can still be categorized as phrases of politeness and general communication.

Metamotifs did prove to be beneficial in classification tasks, but we found them extremely difficult to interpret. Each metamotif is a unique combination of two motifs, which may each be comprised of up to four semantic frames. These combinations of up to eights frames are very rare. Even searching for common metamotifs in the whole corpus yields only a small selection. We see that metamotifs with higher number of occurrences are almost always specific combinations of common phrases or forms of addressing:

- Metamotif: (Statement - Request) + (Text creation - Collaboration - Collaboration)
  Example: Ich eröffne die Aussprache. Zunächst erteile ich das Wort dem Kollegen Gregor Gysi für die Fraktion der Linken.
  (I open the debate. First, I give the floor to fellow Gregor Gysi of the Left Party.)

### Motifs of Political Parties

In the documents of the German Manifesto data set, we investigated the most prominent motifs for each individual party. These motifs were, in fact, very distinctive between the parties, and captured characteristic attitudes and core topics. Below, we show a selection of prominent motifs for some parties, and a summarizing explanation to its occurrences.

**AfD**

- Motif: Judgment communication - Judgment communication - Judgment communication - Judgment communication
  Explanation: Criticism of status quo, describing threat

- Motif: Number - Being employed - Change position on a scale

  Explanation: Fear of rising unemployment

**B90/Die Grünen**

- Motif: People - People - People - People

  Explanation: Listing groups of people, different ages and sexes
- Motif: Political locales - Political locales

  Explanation: Development aid, funding of countryside regions

**CDU/CSU**

- Motif: Collaboration - Collaboration - Collaboration - Collaboration

  Explanation: Networking, synergies with other states
- Motif: Employing - People - Leadership - Undergo change

  Explanation: Demographic changes, pension policy

**Die LINKE**

- Motif: Being employed - Employing - Employing

  Explanation: Rights of employees, unions
- Motif: People - People - People - People

  Explanation: Listing groups of people, different ages and sexes

**FDP**

- Motif: Commerce - Commerce

  Explanation: Personal wealth, taxes
- Motif: Request - Judgment communication

  Explanation: General factional requests

**SPD**

- Motif: Collaboration - Collaboration - Collaboration - Collaboration

  Explanation: Networking, connections, education, research
- Motif: People - People - People - People

  Explanation: Listing groups of people, different ages and sexes

We notice a number of observations in our results. Party issues seem to translate well into prominent motifs. For instance, the core issues of "Die LINKE" include employee rights and anti-discrimination, which are properly represented in the motifs. Among the top motifs, there are only a few cases where the combination of frames reveals a understandable "narrative structure" that directly defines a position or statement. Examples are "Number - Being employed - Change position on a scale" of AfD, or "Employing - People - Leadership - Undergo change" of CDU/CSU. Overall, longer motifs tend to be homogeneous, containing multiple occurrences of the same or similar semantic frames. We observe two distinct reasons:

a) The motif "People - People - People - People" encompasses enumerations of society groups, e.g women and men, young and old. These enumerations are commonly used in the context of equality and anti-discrimination.
b) Other homogeneous motifs are mostly connected to very specific policy issues, like clusters of the "Recht" frame for rules of law, or "Employing" for employees' rights.

For individual frames, we notice that frames link to issues, but can be used for opposite issues across parties. For instance, conservative parties use "Being employed" in statements about effective economy, whereas left wing parties stress employees' rights. The "Judgment communication" frame tends to be used in statements of parties' core values. This frame includes words of high emotional degree, like "praise", "acclaim" or "protest", which strengthens this observation.

## 7.9 Conclusion and Outlook

In previous experiments, we have shown that motif analysis can help in prediction tasks, but also give insights about the underlying reasons for the predictions. Here, we explored the hypothesis that more complex types of motifs can improve the performance in classification tasks. We demonstrated this potential in a party classification task on political speeches. There, metamotifs revealed higher predictive power than simpler motifs and methods. This demonstrates that metamotif analysis has high potential, although the adaptability and scalability of this method to different tasks and data types is still an open question.

Metamotifs can find patterns within patterns, which is useful for generalizing over motifs as features in machine learning scenarios. As expected, these advantages do not project well into qualitative analysis. We suspect that in our case, semantic frames might not generalize **enough**, and even more abstract layers may enable motifs and metamotifs to capture patterns in general lines of arguments or persuasion, rather than patterns based on common phrases. Still, the combination of motif analysis with semantic frames yielded motifs that projected the core issues of political parties reasonably well.

# 8 Summarizing Conclusion

Graph- and motif-based approaches have proven to be powerful tools in classic prediction and classification tasks. But motif analysis can go further, derive new knowledge from the interpretation of discovered motifs. In this thesis, we investigated the possibilities of extended motif-based approaches on textual data, and discussed three main research questions.

First, we applied motif analysis to assess text quality. In the context of the open encyclopedia Wikipedia, we demonstrated a way to extract linguistic patterns that indicate high or low article quality. In addition, we were also able to interpret the results. Motifs that appear in low quality text imply repetitive writing style, whereas forms of cohesive explanation can be found in motifs of high quality text. We confirmed the power of motif analysis in another experiment on encyclopedic data, but on a completely different level. We used different online communities and compared the behavior of users in the collaborative writing process. There, we found motifs in the interaction between specific user groups that have beneficial or detrimental effect on the overall community success. In particular, user groups that specialize on small corrections and other surface level improvements tend to work together in a fruitful way.

Second, we looked at the evolution of motifs in changing texts, again in the context of Wikipedia, but this time with a special focus on local text changes in the revision process. There, we introduced *metamotifs*, or motifs of motifs. We built these metamotifs around local changes in the text, and defined the concept of metamotif stability. Some (unstable) patterns appeared in the revision process of a document only to be revised

again almost every time. Other (stable) patterns seemed to be a desirable "fixed point" in the revision process, as they tend to be never changed again. The stability showed a strong connection to the size and shape of metamotifs. In general, small metamotifs have higher stability than bigger ones, and fully connected metamotifs have higher stability than those that are loosely connected. We also connected motif stability to text quality, and observed that higher quality articles contain more stable motifs.

Finally, we quantified the power of metamotifs compared to simpler methods in a final experiment on political data. We used semantic frames as a layer of abstraction. These frames connect a specific word to a more generic, prototypical situation. For instance, the frame "Delivery" includes words that indicate a deliverer, objects that are handed over, and a recipient. Using these frames, we extracted metamotifs in political speeches of US presidency candidates and German Bundestag debates. These metamotifs constantly showed higher performance in prediction tasks compared to regular motifs and simple frame counts. Additionally, the combination of motif analysis and semantic frames yielded understandable motifs that plausibly reflect the political alignment of German parties. Although metamotifs have proven to be useful features, they are very difficult to interpret, due to rarity and size.

We demonstrated several examples for successful application of motif analysis on textual data, and developed a number of extensions. Motif stability is a new concept in dynamic graphs, and the scope and features of this property is an open research question. There are many possibilities for future applications. Social online networks like Facebook and Twitter are dynamic by nature, and an inspection of the group dynamics with changing motifs might be fruitful. Other interesting data sets can be found in real world networks, like trade or publication networks. Studies in this direction could also

focus on the connection between motif stability and established metrics, like PageRank [65] or node centrality.

We have now shown examples of two different metamotif approaches in textual data, but the general transferability of this method can only be shown by additional research. Motifs and metamotifs can be compared to other features in many machine learning tasks, including, but not limited to, the field of NLP. Combining metamotifs with other features might also open new research questions. In recent years, the most dominant methods in machine learning are based on various neural network architectures. Since motif signatures, by design, can be seen as vector representations of the source data, they can be tested as substitutes or supplements to other vector based input in existing neural network architectures.

# 9 Acknowledgements

I want to express my sincere gratitude to my supervisor Professor Karsten Weihe from the Algorithmics Group at the Department of Computer Science, TU Darmstadt. His lectures on graph algorithms and optimization techniques during my studies as a master's student really piqued my interest in these topics. He also guided me during my master thesis on coreference resolution, which started my interdisciplinary research on natural language topics. In my three years as a PhD student in the AIPHES research training group, he always took his time to listen to my ideas, give insightful advice, and ask the right questions.

I also want to thank my committee members, Professor Iryna Gurevych, Professor Johannes Fürnkranz, Professor Christian Reuter und Professor Michael Goesele. Special thanks go to my external supervisor and final member of my committee, Professor Matthias Müller-Hannemann from the Martin-Luther-Universität in Halle. Especially at the beginning of my PhD research, his advice and ideas really helped me to specify and focus on target research questions.

For good results, you not only need ideas and effort, but also a good and supporting working environment. Therefore, I would like to thank the whole AIPHES research training group, that provided both necessary space and infrastructure, but also helpful training and workshops, interesting talks and opportunities to exchange ideas with leading members of the scientific community, and a stage to present and discuss my current work. Of course, I want to thank my awesome colleagues from AIPHES, the Algorithmics group, UKP and all other people I had the pleasure to work with, for their inspiring

ideas, motivating comments, and also deep discussions at our philosophical Friday coffee meetings. I would like to mention and give thanks to a number of colleagues that I closely worked with on certain projects. In this context, thank you, Johannes Daxenberger, for our great cooperation on Wikipedia and Wikia data, Teresa Botschen, for your effort and dedication in my experiments on semantic frames, and Pascal Jürgens together with his supervisor, Professor Birgit Stark, for the fruitful interdisciplinary collaboration.

Finally, I want to thank my friends and my family for their constant support. In particular, I thank my wonderful wife Michaela, for her belief in my work and skills, and for her help and support in every way.

# Bibliography


[1] Hala AbouAssi, Yu Chen Lin, Israel Serrano, Carlos Gonzales, and Masad J. Damha. Probing Synergistic Effects of DNA Methylation and 2ˊBeta-Fluorination on i-Motif Stability. *Chemistry – A European Journal*, 24(2):471–477. (Cited on page 73.)

[2] Charu C Aggarwal. *Social Network Data Analytics*. Kluwer, 2011. (Cited on page 11.)

[3] Charu C. Aggarwal and Haixun Wang. Managing and Mining Graph Data. *Database*, 40:487–513, 2010. (Cited on page 11.)

[4] Younis Al Rozz and Ronaldo Menezes. Author Attribution Using Network Motifs. In Sean Cornelius, Kate Coronges, Bruno Gonçalves, Roberta Sinatra, and Alessandro Vespignani, editors, *Complex Networks IX*, pages 199–207, Cham, 2018. Springer International Publishing. (Cited on page 85.)

[5] Aris Anagnostopoulos, Luca Becchetti, Carlos Castillo, Aristides Gionis, and Stefano Leonardi. Online Team Formation in Social Networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 839–848, Lyon, France, 2012. (Cited on pages 68 and 69.)

[6] Judd Antin, Coye Cheshire, and Oded Nov. Technology-Mediated Contributions: Editing Behaviors Among New Wikipedians. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 373–382, Seattle, WA, USA, 2012. (Cited on page 51.)

[7] Lucas Antiqueira, M das Gracas V Nunes, ON Oliveira Jr, and L da F Costa. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, 373:811–820, 2007. (Cited on page 19.)

[8] Ofer Arazy, Oded Nov, and Nov Oded. Determinants of Wikipedia Quality: the Roles of Global and Local Contribution Inequality. In *Proceedings of the ACM 2010*

*Conference on Computer Supported Cooperative Work*, pages 233–236, Savannah, GA, USA, 2010. (Cited on page 44.)

[9] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. Functional Roles and Career Paths in Wikipedia. In *Proceedings of the 18th Conference on Computer Supported Cooperative Work and Social Computing*, pages 1092–1105, Vancouver, BC, Canada, 2015. (Cited on page 47.)

[10] Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. Turbulent Stability of Emergent Roles: The Dualistic Nature of Self-Organizing Knowledge Co-Production. *Information Systems Research. Articles in Advance*, 2016. (Cited on pages 44, 48, 51, 52, 53, 55, 56, 66, and 67.)

[11] Ofer Arazy, Hila Lifshitz-Assaf, Oded Nov, Johannes Daxenberger, Martina Balestra, and Coye Cheshire. On the "How" and "Why" of Emergent Role Behaviors in Wikipedia. In *Proceedings of the 20th Conference on Computer-Supported Cooperative Work and Social Computing*, page to appear, Portland, OR, USA, 2017. (Cited on page 69.)

[12] Thomas Arnold and Karsten Weihe. Network Motifs May Improve Quality Assessment of Text Documents. In *Proceedings of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing*, pages 20–28, San Diego, CA, USA, 2016. (Cited on page 18.)

[13] Thomas Arnold, Johannes Daxenberger, Karsten Weihe, and Iryna Gurevych. Is Interaction More Important Than Individual Performance? A Study of Motifs in Wikia. In *Proceedings of the 26th International Conference Companion on World Wide Web*, WWW '17 Companion, pages 1609–1617. International World Wide Web Conferences Steering Committee, April 2017. (Cited on page 44.)

[14] David Arthur and Sergei Vassilvitskii. K-means++: The Advantages of Careful Seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, New Orleans, LA, USA, 2007. (Cited on page 55.)

[15] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998. (Cited on pages 12, 83, and 89.)

[16] Johannes Berg and Michael Lässig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41):14689–14694, 2004. (Cited on page 13.)

[17] Adam Bermingham and Alan Smeaton. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, 2011. (Cited on page 85.)

[18] Chris Biemann, Stefanie Roos, and Karsten Weihe. Quantifying semantics using complex network analysis. In *Proceedings of COLING 2012, the 26th International Conference on Computational Linguistics*, 2012. (Cited on pages 13, 15, and 17.)

[19] Chris Biemann, Lachezar Krumov, Stefanie Roos, and Karsten Weihe. *Network Motifs Are a Powerful Tool for Semantic Distinction*, pages 83–105. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016. (Cited on page 15.)

[20] Norman Biggs, E Keith Lloyd, and Robin J Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1976. (Cited on page 11.)

[21] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. (Cited on page 62.)

[22] Joshua E Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*, pages 1095–1096. ACM, 2008. (Cited on pages 17 and 23.)

[23] Hans Christian Boas. Bilingual FrameNet dictionaries for machine translation. In *LREC*, 2002. (Cited on page 84.)

[24] Leo Born, Mohsen Mesgar, and Michael Strube. Using a Graph-based Coherence Model in Document-Level Machine Translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 26–35, 2017. (Cited on page 15.)

[25] Teresa Botschen, Hatem Mousselly-Sergieh, and Iryna Gurevych. Prediction of Frame-to-Frame Relations in the FrameNet Hierarchy with Frame Embeddings. In *Proceedings of th 2nd Workshop on Representation Learning for NLP (RepL4NLP, held*

*in conjunction with ACL 2017)*, pages 146–156, August 2017. (Cited on pages 85 and 89.)

[26] D. Braha and Yaneer Bar-Yam. *Time-Dependent Complex Networks: Dynamic Centrality, Dynamic Motifs, and Cycles of Social Interactions*, pages 39–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. (Cited on page 73.)

[27] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network Analysis of Collaboration Structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pages 731–740, Madrid, Spain, 2009. (Cited on pages 45, 46, 48, 51, and 58.)

[28] Eric Brill and Robert C Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293. Association for Computational Linguistics, 2000. (Cited on page 19.)

[29] Cody Buntain and Jennifer Golbeck. Identifying Social Roles in Reddit Using Network Structure. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 615–620, Seoul, Republic of Korea, 2014. (Cited on page 47.)

[30] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 192–199. IEEE, 2011. (Cited on page 85.)

[31] D. J. Cook and L. B. Holder. *MIning Graph Data*. Wiley, 2006. (Cited on page 11.)

[32] Johannes Daxenberger and Iryna Gurevych. A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 711–726, Mumbai, India, 2012. (Cited on pages 47 and 68.)

[33] Johannes Daxenberger and Iryna Gurevych. Automatically Classifying Edit Categories in Wikipedia Revisions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, WA, USA, 2013. (Cited on page 51.)

[34] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, 2003. (Cited on page 20.)

[35] Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 537–544. Association for Computational Linguistics, 2003. (Cited on pages 84 and 89.)

[36] Nicole Ernst, Rinaldo Kühne, and Werner Wirth. Effects of Message Repetition and Negativity on Credibility Judgments and Political Attitudes. *International Journal of Communication*, 11:21, 2017. (Cited on page 100.)

[37] Oliver Ferschke. *The quality of content in open online collaboration platforms: Approaches to NLP-supported information quality management in Wikipedia*. PhD thesis, Technische Universität, 2014. (Cited on page 19.)

[38] Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 97–102, Portland, OR, USA, 2011. (Cited on page 50.)

[39] Charles J Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976. (Cited on page 89.)

[40] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010. (Cited on page 11.)

[41] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, 31(6):649–679, 2013. (Cited on page 85.)

[42] Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. Overview of the 2003 KDD Cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003. (Cited on page 38.)

[43] Ana-Maria Giuglea and Alessandro Moschitti. Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the 21st International Conference on Com-*

*putational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 929–936. Association for Computational Linguistics, 2006. (Cited on page 84.)

[44] Kalanit Grill-Spector, Richard Henson, and Alex Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006. (Cited on page 30.)

[45] Ruben Grothwinkel. A Graphical User Interface for a Motif Analysis Toolkit. Bachelor's thesis, TU Darmstadt, 2017. (Cited on page 31.)

[46] Jörg Hau, Michael Muller, and Marco Pagni. HitKeeper, a generic software package for hit list management. *Source Code for Biology and Medicine*, 2(1):2, Mar 2007. (Cited on page 73.)

[47] David Jurgens and Tsai-ching Lu. Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia. In *Proceedings of the 6th International Conference on Weblogs and Social Media*, pages 162–169, Dublin, Ireland, 2012. (Cited on pages 46, 47, 48, 60, and 69.)

[48] Betty R. Kirkwood and Jonathan A. C. Sterne. *Essentials of Medical Statistics.* Blackwell Scientific Publications, 1988. (Cited on page 60.)

[49] Aniket Kittur, Ed H. Chi, Bryan Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the Few vs. Wisdom of the Crowd: Wikipedia and the rise of the bourgeoisie. In *Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems*, San Jose, CA, USA, 2007. (Cited on page 44.)

[50] Klaus Krippendorff. *Die semantische Wende: eine neue Grundlage für Design*. Walter de Gruyter, 2012. (Cited on page 83.)

[51] Lachezar Krumov, Christoph Fretter, Matthias Müller-Hannemann, Karsten Weihe, and Marc-Thorsten Hütt. Motifs in co-authorship networks and their relation to the impact of scientific publications. *European Physical Journal B*, 84(4):535–540, 2011. (Cited on page 15.)

[52] David Laniado and Riccardo Tasso. Co-authorship 2.0: Patterns of Collaboration in Wikipedia. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 201–210, Eindhoven, The Netherlands, 2011. (Cited on page 45.)

[53] Mirella Lapata. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484, 2006. (Cited on page 19.)

[54] Jun Liu and Sudha Ram. Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality. *ACM Transactions on Management Information Systems*, 2(2):11:1–11:23, 2011. (Cited on pages 44, 47, and 55.)

[55] Annie Louis and Ani Nenkova. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics*, 1:341–352, 2013. (Cited on page 19.)

[56] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. (Cited on page 37.)

[57] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. (Cited on page 90.)

[58] Mohsen Mesgar and Michael Strube. Graph-based Coherence Modeling For Assessing Readability. *Lexical and Computational Semantics (* SEM 2015)*, page 309, 2015. (Cited on pages 15 and 17.)

[59] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. (Cited on page 12.)

[60] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663): 1538–1542, March 2004. (Cited on page 14.)

[61] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002. (Cited on pages 14 and 59.)

[62] Baharan Mirzasoleiman and Mahdi Jalili. Failure tolerance of motif structure in biological networks. *PLoS One*, 6(5):e20512, 2011. (Cited on page 13.)

[63] Javier Onrubia and Anna Engel. Strategies for Collaborative Writing and Phases of Knowledge Construction in CSCL Environments. *Computers & Education*, 53(4): 1256–1265, dec 2009. (Cited on page 44.)

[64] Ekaterina Ovchinnikova, Laure Vieu, Alessandro Oltramari, Stefano Borgo, and Theodore Alexandrov. Data-driven and ontological analysis of FrameNet for natural language reasoning. In *LREC*, 2010. (Cited on page 84.)

[65] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999. (Cited on page 107.)

[66] Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. Learning to Score System Summaries for Better Content Selection Evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, 2017. (Cited on page 85.)

[67] Matias Piipari, Thomas A. Down, and Tim JP Hubbard. Metamotifs - a generative model for building families of nucleotide position weight matrices. *BMC Bioinformatics*, 11(1):348, Jun 2010. (Cited on page 73.)

[68] Reid Priedhorsky, Jilin Chen, Shyong K Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, Destroying, and Restoring Value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, pages 259–268, Sanibel Island, FL, USA, 2007. (Cited on page 44.)

[69] The Opte Project. `http://www.opte.org/the-internet/`. Accessed: 11.04.2018. (Cited on page 2.)

[70] J. Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014. (Cited on page 23.)

[71] Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. *FrameNet II: Extended theory and practice*. Institut für Deutsche Sprache, Bibliothek, 2016. (Cited on page 84.)

[72] Mrinmaya Sachan, Danish Contractor, Tanveer A Faruquie, and L Venkata Subramaniam. Using Content and Interactions for Discovering Communities in Social Networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 331–340, Lyon, France, 2012. (Cited on page 46.)

[73] Falk Schreiber and Henning Schwöbbermeyer. MAVisto: a tool for the exploration of network motifs. *Bioinformatics*, 21(17):3572–3574, 2005. (Cited on page 13.)

[74] Falk Schreiber and Henning Schwöbbermeyer. *Statistical and Evolutionary Analysis of Biological Network Data*, chapter Motifs in biological networks, pages 45–64. Imperial College Press/World Scientific, 2010. (Cited on page 14.)

[75] Karin Kipper Schuler. VerbNet: A broad-coverage, comprehensive verb lexicon. 2005. (Cited on page 12.)

[76] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–68, 2002. (Cited on page 14.)

[77] Lei Shi and Rada Mihalcea. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 100–111, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. (Cited on page 85.)

[78] Deborah Tannen. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*, volume 26. Cambridge University Press, 2007. (Cited on page 30.)

[79] Joel R. Tetreault and Martin Chodorow. The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 865–872, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. (Cited on page 19.)

[80] Ngoc Tam L Tran, Luke DeLuccia, Aidan F McDonald, and Chun-Hsi Huang. Cross-disciplinary detection and analysis of network motifs. *Bioinformatics and Biology insights*, 9:49, 2015. (Cited on page 15.)

[81] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-Labelsets for Multi-Label Classification. *IEEE Trans. on Knowledge and Data Engineering*, 23 (7):1079–1089, 2011. (Cited on page 52.)

[82] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, Vienna, Austria, 2004. (Cited on page 45.)

[83] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. Wisdom in the Social Crowd: An Analysis of Quora. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1341–1352, Rio de Janeiro, Brazil, 2013. (Cited on pages 44 and 46.)

[84] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding Social Roles in Wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129, Seattle, WA, USA, 2011. (Cited on page 47.)

[85] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and Quality in the Wikipedia. In *Proceedings of the International Symposium on Wikis and Open Collaboration*, pages 157–164, Montreal, Canada, 2007. (Cited on page 44.)

[86] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User Interactions in Social Networks and Their Implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*, pages 205–218, Nuremberg, Germany, 2009. (Cited on pages 44 and 46.)

[87] Ian H Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with Java implementations. 1999. (Cited on pages 23 and 96.)

[88] William A Woods. What's in a link: Foundations for semantic networks. In *Representation and understanding*, pages 35–82. Elsevier, 1975. (Cited on page 12.)

[89] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. Characterizing Wikipedia Pages using Edit Network Motif Profiles. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 45–52, Glasgow, Scotland, UK, 2011. (Cited on page 48.)

[90] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. Edit Categories and Editor Role Identification in Wikipedia. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1295–1299, Portorož, Slovenia, 2016. (Cited on page 47.)

[91] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. Who Did What: Editor Role Identification in Wikipedia. In *Proceedings of the 10th international AAAI Conference on Web and Social Media*, pages 446–455, Cologne, Germany, 2016. (Cited on pages 47 and 48.)

[92] Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. Revealing the hidden language of complex networks. *Scientific reports*, 4:4547, 2014. (Cited on page 73.)

[93] Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. The Next Step for Multi-Document Summarization: A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1535–1545, 2016. (Cited on page 37.)

# List of Figures

# List of Tables